UNIVERSITÉ CATHOLIQUE DE LOUVAIN
INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES,
ELECTRONICS AND APPLIED MATHEMATICS
DEPARTMENT OF APPLIED MATHEMATICS

Data Science for Modeling Opinion Dynamics on Social Media

Corentin Vande Kerckhove

Thesis submitted in partial fulfillment of the requirements for the Ph.D. Degree in Engineering Sciences

Thesis Committee

Advisors -	Vincent Blondel	*Université catholique de Louvain*
	Julien Hendrickx	*Université catholique de Louvain*
Jury -	Jason Rentfrow	*University of Cambridge*
	Yannick Dufresne	*Université Laval*
	Jean-Charles Delvenne	*Université catholique de Louvain*
Chairman -	Philippe Lefèvre	*Université catholique de Louvain*

Louvain-la-Neuve, September 11, 2017.

Acknowledgments

This long journey started innocuously with a sunny afternoon of May 2013, in the French Ardennes, during a scientific retreat in the fabulous setting of the Moulin de Daverdisse. Outside, groups of professors and researchers were gathered to tackle open mathematical problems that I have to admit seemed unsolvable to me at that time. Before that event, I actually never really intended to start a research carrier, because of the misguided image I had about the academic world. I instead realized later that I never learned as much as I did during those four years of research.

On that same day, I met my two supervisors—Vincent Blondel and Julien Hendrickx—whom I would like to thank for having convinced me to join their research group and for their careful guidance during all these years in the Mathematical Engineering department. In particular, I am grateful to Vincent for the numerous opportunities he offered me, opportunities that allowed me to work with many successful research teams. It is partly thanks to his support that I was fortunate enough to be funded by the FNRS-FRIA, whom I would like to thank too. I also would probably never have been able to dive so deeply into my projects without the great contributions and ideas of Julien. His ability to put forward novel research directions and to deal quickly with the potential weaknesses of a project makes him without any doubt an outstanding advisor. This thesis has also been reviewed by a dissertation committee, composed of Pr Jason Rentfrow, Pr. Yannick Dufresne, Pr. Jean-Charles Delvenne and Pr. Philippe Lefèvre. I would like to thank Jason and Yannick for sharing their respective expertise in Social Psychology and Political Science which have been crucial regarding the multidisciplinary path this thesis has followed. Jean-Charles, thank you for your availability every time I felt lost in my project—and for having shown me the exquisite links between mathematics and magic tricks. Finally, I am also grateful to Philippe who chaired the jury with a great professionalism during the dissertation process and who provide me an insightful feedback on my work.

During these four years, I was very fortunate to share my office with Matthew Philippe, probably one of the most talented and promising researchers I have ever met—Matthew, remember, you are a scholar and a gentleman! Amongst the many collaborators who decided to visit us, I would like to take the time to mention three researchers whose contributions substantially impact the present document. In that sense, I am grateful to Samuel Martin for having trusted me since the beginning of my PhD and for having guided me through my first publication. Also, I would hardly forget the year that I spent in Québec with Mickael Temporão whom I really consider as a true friend. I would like to thank him for having taught me an innumerable number of tips, for his constant energy and for his warm welcome. During my stay, you really showed us that collaborations between social scientists and engineers are really beneficial for the research community. I also want to take this opportunity to thank the Vox Pop Labs team—especially Clifton Van Der Linden—for the trust they have placed in me during my first collaboration with a private company. Following my return from Canada, I finally had the chance to collaborate with Balazs Gerencsér, whom I would like to thank for the insightful discussions that we had during all your comings in Belgium.

I very much enjoyed the daily life at Euler and will never forget this department with a great sense of humor and led by amazing members. From the delicious pies to the foosball games, I would also like to mention a series of joyful people : Romain and Adeline, who were very welcoming when I first arrived. Isabelle, Nathalie and Marie-Christine, who have done a remarkable job and helped me countless times. Pierre-Yves, for bringing so much energy to this department by sharing crazy stories to the community—those will for sure be remembered by Etienne and François W., whom I would also like to thank here. Luc, for the enlightening and delightful conversations we had in the numerous coffee shops of Brussels. Allan, David, François G., Adrien, Pierre-Alexandre Beaufort and all the others that made Euler such a pleasant and vibrant work environment.

Enfin, je souhaiterais remercier tous ceux qui ont contribué à cette thèse de près ou de loin, que ce soit par leur soutien, leurs idées ou par les nombreux projets que nous avons partagés. Par ordre chronologique, je remercie mes colocataires de la Ramée 18 de m'avoir autant diverti après le travail et à ceux de la Ramée 7 d'avoir pris la relève l'année d'après. Je suis reconnaissant envers les membres de l'ASCII d'avoir eu confiance dans la construction d'un vélo des 24 heures pour notre institut, et j'en remercie l'actuel président d'avoir cru en la création de notre salle de détente, un lieu particulièrement propice au bouillonnement

scientifique. Je remercie également les membres de l'IEEE Student Branch et je les encourage à continuer leur investissement dans ces projets qui façonnent nos instituts de recherche. Un grand merci à mes 25x11 étudiants d'être parvenus à lire les nombreuses formules que j'avais gribouillées devant eux, ainsi qu'à mon équipe de Markstrat d'avoir toléré mes absences tout au long de notre master en gestion. Je n'oublie évidemment pas mes collègues politologues québécois (Justin, Audrey, Gab et tant d'autres) que je remercie pour leur accueil ainsi que d'avoir accepté la présence incongrue d'un ingénieur dans leur bureau pendant une année. Merci à Fanny n°2 d'avoir pris la peine de relire avec tellement d'attention certaines parties de ma thèse.
Enfin, au terme de cette longue traversée, je remercie ma conjointe, Anne-Sophie, d'avoir toléré mes nombreux moments de réflexion, parfois jusqu'à tard le soir. Je te suis vraiment reconnaissant de m'avoir tellement soutenu tout au long de cette épreuve. J'aimerais consacrer ces dernières lignes de remerciement à ma famille, et particulièrement à mes parents sans qui je n'aurais eu la chance de découvrir la richesse du milieu universitaire. Je ne saurais que trop vous remercier.

Contents

Notations

Opinions and Ideologies

x — Opinions of agents

$\bar{x}$ — Average opinion of the group

$\tilde{x}$ — Predicted opinions of agents

X — Measured opinions of agents

ϕ — Ideologies of agents

$\hat{\phi}_{net}$ — Network estimates of agents' ideologies

$\hat{\phi}_{txt}$ — Textual estimates of agents' ideologies

T — True value

t — Time variable

Sample sizes

n — Number of agents (citizens) in the network

m — Number of agents (elites) in the network

W — Number of words in the political dictionary

Variables and Parameters

η (resp. $\bar{\eta}$) — Intrinsic noise after the second (resp. third) round

a_{ij} — Weight that an agent i put on the judgment of an agent j

α_i — Influenceability of an agent i towards the average

γ — Power-law exponent

θ Factor loadings

p Popularity parameters

s Susceptibility parameters

Matrices

A Matrix of advice taking weights

Z Term-Document matrix

Acronyms

MSE Mean Square Error
MAE Mean Absolute Error
RMSE Root Mean Square Error
EM Expectation Maximization
TDM Term-Document matrix
AUC Area Under Curve
ML Maximum Likelihood
MK Markovian process
IT Independent Threshold Model
IP Independent Power-law Model

Chapter 1

Introduction

Individuals are more and more connected. This is particularly noticeable on the online world. Recent technologies such as the Web 2.0 with user-generated content have allowed individual actions to be combined through social interactions at a scale that was previously unreachable. Since the last decade, the arrival of social media has established new practices which have changed the way humans interact. Nowadays, anyone can access personal information by simply following a subset of relevant users with a social media account. Users can take part in a debate from their smartphones with a simple like or share. They can even express their own feelings on different topics by posting a message. Some social media such as Twitter make their content public – meaning that everything shared on the platform is freely available – while others are more prohibitive such as Facebook. In 2017, they are over 317 million monthly active users on Twitter [Omnicore (2017)] and no less than 1.94 billion monthly active Facebook accounts [Zephoria (2017)]. The availability of online data has led to a recent surge in trying to understand how social influence impact human behavior.

Many individual judgments are mediated by observing others' opinions. This is true for buying products, voting for a political party or choosing to donate blood. Although decision outcomes are often tied to an objective best choice, outcomes can hardly be fully inferred from this supposedly best choice. For instance, predicting the popularity of songs in a cultural market requires more than just knowing the actual song quality [Salganik, Dodds, and Watts (2006)]. The decision outcome is rather determined by the social influence process at work [Lorenz, Rauhut, Schweitzer et al. (2011)]. It is also known that groups can outperform the best of its members [Clearwater, Huberman, and Hogg (1991)]. Collective intelligence allows reaching goals that no single individual could

have achieved before. Wikipedia, as the world largest encyclopedia, is a good example. A first step towards designing collective intelligence mechanisms is to understand how individuals' past behavior and microlevel attributes (such as personalities) affect the group outcome.

Detecting influent actors and measuring accurately people's opinion is a major challenge for scientists, especially in the field of political science. Attitudes of citizens towards political issues are one of the major factors that explain voting behaviour [Conover and Feldman (1981)]. To date, most of the research in this area is done by designing and conducting surveys that target political elites, ordinary citizens or a combination of both. Nonetheless, repeatedly obtaining survey data on a large scale, across countries, and over time is often prohibitively costly [Bond and Messing (2015); Slapin and Proksch (2008)]. Recently, alternative methods are emerging, focusing on the extraction of ideological positions from social media data [Barberá (2015); Bond and Messing (2015)]. This raises a broad set of issues concerning the validity of the estimates and whether their effects remain consistent throughout these methods. Validation of such measures for political parties' and candidates' ideologies is eased by the public nature of these elites' political stances. But it is much harder to validate ordinary citizens' extracted ideologies. Some of the data generated by social media can even be considered in some extent as richer than conventional surveys.

All of the above highlights a set of relevant questions concerning the study of online opinions. For instance, how precisely can we predict the variations of judgments resulting from social influence, even when the information is scarce? Does social influence promote or undermine the performance of a group? Are we able to predict the voting behavior of citizens from their online posts and comments? In this thesis, we aim at understanding how opinions are expressed on social media, and whether their evolution can be anticipated by taking the interaction between agents into account. Although our contributions spread over various research topics, Figure 1.1 attempts to summarize the structure of the thesis around its different chapters.

Chapter 2 : The Study of Opinions on Social Networks

This preliminary chapter introduces various notions related to the analysis of online opinions. We start by discussing the representation of opinions in space and time, and detail a well-known effect – the wisdom of the crowd – that occurs when aggregating independent answers. Opinions of agents are affected by influence networks, which are represented in this work by graph

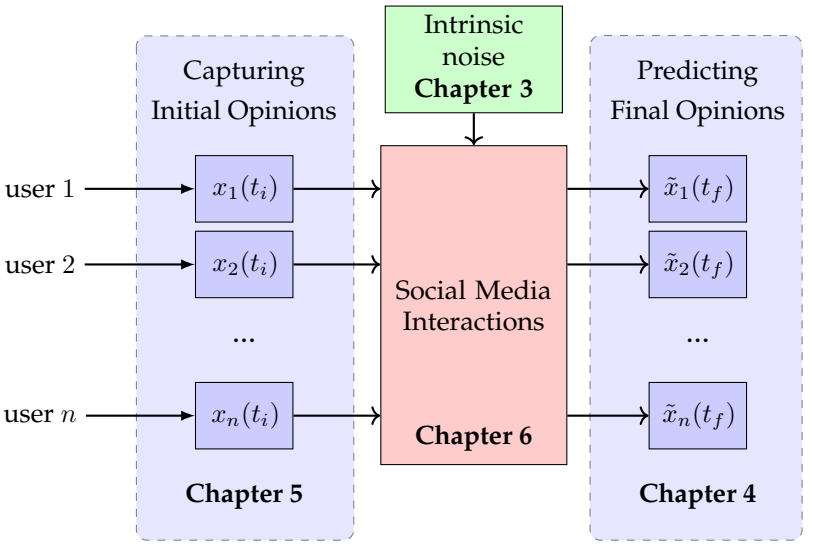


Fig. 1.1 Global overview of the thesis.

theory, a mathematical tool that models pairwise interactions. The chapter ends with a discussion on the use of crowdsourcing platforms for experimenting and analyzing human influence.

Chapter 3 : The Unpredictability of Human Judgment Revision

Judgment revision includes unpredictable variations that limit the potential for prediction. In Chapter 3, we propose a measure of unpredictability based on online experiments. When people are asked to give their opinion twice with some delay in between, scholars observe that participants often express different answers. We provide a measure of what we will call the *intrinsic noise* in judgment revision. Estimating the intrinsic unpredictability allows to assess the prediction efficiency of opinion dynamic models. It provides a limit beyond which no one can expect to improve predictions.

The results presented in this chapter were derived using crowdsourcing experiments, where each participant repeated several times estimation tasks in very similar conditions. Our results have been reviewed and accepted for an oral presentation at the *2nd Annual International Conference on Computational Social Science* (2016 - Evanston). It has been published altogether with the content of Chapter 4 in one single paper [Vande Kerckhove, Martin, Gend et al. (2016)].

Chapter 4 : Analyzing Opinion Revision with Consensus models

Although models of opinion dynamics have been widely studied for decades by sociologists from a theoretical standpoint, the predictive power of a quantitative model of opinion dynamics needs to be tested against a real dataset. Chapter 4 shows how the so-called *time varying consensus model* can be used to predict opinion evolution in online collective behavior.

By contrast to previous research on the topic, the model is validated on data that did not serve to calibrate it. This avoids favoring more complex models over simpler ones and prevents overfitting. Our approach casts individuals' behaviors according to their *influenceability*, the factor quantifying to what extent one takes external opinions into account. The analysis discuss multiple prediction scenarios according to different levels of prior knowledge on the participants. The prediction accuracy depends on the participants' past behavior. Several situations reflecting data availability are compared. When the data is scarce, the data from previous participants is used to predict how a new participant will behave. More than two thirds of the prediction errors are found to occur due to unpredictability of the human judgment revision process rather than to model imperfection.

We further expressed some practical implications, including discussions about group performance and potential relations between influenceability and personality traits. This work has been accepted for an oral presentation at the first edition of the *Annual International Conference on Computational Social Science*, in 2015. As already mentioned, it has later been published altogether with the content of Chapter 3 [Vande Kerckhove, Martin, Gend et al. (2016)].

Chapter 5 : Uncovering Political Ideologies from Digital Traces

The previous in vitro experiments rely on the knowledge of the participants' initial opinions. How are these opinions expressed and captured on social media? Chapter 5 deals with the particular case of political ideologies. Words matter when it comes to politics and social media platforms are increasingly becoming the medium of politics. Yet, existing ideological scaling methods fail when applied to short informal textual content, which is abundant on platforms such as Twitter. Our main contribution proposes a viable approach to measure individual-level political positions based on the contents that are posted on social media outlets. This method allows us to position political parties, politicians, and citizens who are active on social media on a comparable ideological dimension.

The proposed method is tested on two types of political actors (elites and citizens), in three political contexts (Canada, New Zealand and Québec[1]) and two different languages (French and English). These results are shown to be valid and consistent by using a unique large-scale survey. The survey allows us to bridge the gap between classic survey data and social media data. We show that the proposed method is able to accurately capture ideologies, which are also shown to provide predictive value on vote choice. Our results have been reviewed and accepted to three peer-reviewed conferences[2]. A journal paper has been submitted to the peer-reviewed journal *Political Analysis* [Temporão, Vande Kerckhove, Hendrickx et al. (2016)].

Chapter 6 : Memory in inter-communication times

In the last chapter, we investigate the arising communication patterns on social media, and in particular the series of events happening for a single user. The distribution of waiting times separating two consecutive events – also denoted by the inter-event distribution – is well known to have a density fitting a power-law function [Aledavood, López, Roberts et al. (2015); Clauset, Shalizi, and Newman (2009); Johansen (2004)]. A debate, however, persists on the nature of an underlying model that explains the observed distribution, and whether the model should incorporate an inter-event dependence.

The purpose of the present chapter is to integrate the dependence of consecutive waiting times into the power-law model. One can observe that social media behavior is characterized by periods of intense activities separated by longer periods of inactivity. We propose an intuitive explanation to understand the observed dependence of subsequent waiting times. This leads us to create a model that incorporates memory. Our contribution is twofold. The first idea consists of separating the short waiting times – out of scope for power-law distributions – from the long ones. The model is further enhanced by introducing a two-state Markovian process to incorporate memory. Both contributions show a significant improvement for modeling commenting events on Twitter and Reddit. This last project is currently under review for publication in Physical Review E.

[1] The thesis also provides a brief analysis of the Belgian case.

[2] The *2016 meeting of the Midwest Political Science Association*, the *2nd Annual International Conference on Computational Social Science* and the *SIAM Workshop on Network Science*

Summary of the major publications

The research results described in this thesis have been presented at various conferences and meetings, and have later submitted or published to the three following scientific journals :

- Vande Kerckhove, C, Martin, S, Gend, P, Rentfrow, PJ, Hendrickx, JM, and Blondel, VD. (2016). Modelling influence and opinion evolution in online collective behavior. *PloS one*, $11(6) : e0157685$.
- Temporão, M, Vande Kerckhove, C, Hendrickx, JM, and Dufresne, Y (2016). On the validity of social media data for public opinion research. Under review in Political Analysis.
- Gerencsér, B, Vande Kerckhove, C, Hendrickx, JM, and Blondel, VD. (2017). Markov modeling of online inter-arrival times. Under review.

Chapter 2

The Study of Opinions on Social Networks

The present chapter introduces various concepts and tools that we consider of particular interest in the analysis of opinion dynamics. In this line of work, we first investigate the notion of opinion and its mathematical formalization. We then deal with collective opinions. In particular, we briefly discuss the assumptions behind the popular *wisdom of the crowd* effect. Thereafter, we consider the possible influences existing between social agents. In this regard, we offer a quick introduction to graph theory that formalizes the study of network structures between interconnected agents. We will close this chapter by guiding the reader through an overview of the ways in which crowdsourcing platforms can be used for experimenting human influence.

2.1 Representation of opinions

An opinion is generally defined as a judgment or view that people hold about objects, agents or more generally any abstract concept (e.g., political issues or services). The vocabulary used to refer to the expresion of a particular feeling varies according to the research area. While political science often favors the term *opinion*, social psychology tends to use the term *attitude* [Bergman (1998)]. Some authors consider that opinions are unemotional beliefs while attitudes involve an emotional assessment [Oskamp and Schultz (2005)]. As an example, they consider that a sentence like « *Barack Obama was a good president* » would make reference to an opinion and that the sentence « *I like Barack Obama* » would rather appear to be an attitude. Other scholars argue that distinguishing attitudes from opinions leads to confusions [McGuire (1969)].

Throughout this work, we share the vision of McGuire (1969) and consider opinions and attitudes as synonymous.

A discussion may be held as to whether the complexity inherent to human opinions can accurately be described by mathematical models. Nowadays, numerous sociologists and engineers continue to support the inefficiency of mathematics to describe the outside world [Abbott (2013)]. White (1943) claims that the most fundamental sociological problems are non-mathematical in nature. In particular, the fuzzy nature of human opinions comes as a real challenge for modeling purposes. However, there is no doubt about the benefits of developping quantitative models to anticipate human behaviors [Kemeny (1972)]. The work of Rand and Rust (2011) is just one of the many examples of how agent-based models have shown a substantial potential to guide marketing decision rules.

Researchers attempt to capture the fuzzy nature of opinions by defining mathematical frameworks. For instance, Wang and Mendel (2016) represent agent opinions by a Gaussian fuzzy set. Others describe the evolution of attitudes in a multidimensional state space [Lorenz (2007)]. In the following, we denote the opinion (or attitude) of an agent i on a particular subject or issue k as x_{ik}. We sometimes omit the index k and write x_i when the particular topic is clearly identified. Opinion dynamics are well adapted to represent opinions evolving in a continuous space. This matches many real world situations in which the opinion represents the inclination, not only for one of two extreme opposite options (such as left and right in politics), but also for any other position in between. Opinions are often represented in a unidimensional space (e.g., $x_{ik} \in \mathbb{R}$) but may lie more generally in a multidimensional space. Alternatively, part of the literature considers models with binary choices or actions. This is the case when people are voting for one candidate, buying a car or going on strikes. Discrete judgments are for instance recorded when participants are asked to express their opinions on Likert scales—for example on a five-level scale, ranging from $x_{ik} = 1$ (strongly disagree) to $x_{ik} = 5$ (strongly agree). The social models developed in this thesis rely on the *rational choice theory* [Becker (2013); Simon (1955)]. In this sense, we make the assumption that individuals express their opinions based on the available alternatives—called the *behavior alternatives*—in order to maximize their own utility, described by *pay-off* functions. The theory further assumes that the individuals consider all pairs of decisions and that their preferences are transitive relations[1].

[1] That is, if decision A is preferred to decision B and if decision B is preferred to decision C, then decision A is preferred to decision C

In some cases, the subject or issue on which the agents are providing an opinion is associated with an ideal value or *ground truth*, allowing to assess the accuracy of an individual. We will see in the following how the individuals' performances can be compared with the aggregate performance of a whole group.

2.2 Aggregating independent judgments

In the presence of a ground truth, the distance between an opinion and the actual value can be seen as a measure of performance. When individual opinions are independent, we observe that the aggregation of these opinions generally leads to a substantial gain in accuracy. This statistical[2] effect, called the *wisdom of the crowd effect*, was first reported by the statistician Francis Galton in 1906. At that time, a weight-judging competition was organized at Plymouth in the southwest of England. More than 700 participants were asked to guess the weight of a fat ox and prizes were given to the players who provided the closest estimations. Francis Galton observed the participants' guesses and discovered that the average opinion was really close to the actual weight of the ox [Galton (1907)].

These experiments of collective intelligence tell us that crowds in most cases have better information than the individuals who are part of it. The wisdom of the crowd effect states that the mean opinion is very often *much* closer to the actual value than the individual opinions. This is illustrated by the two following theorems.

The Diversity Prediction Theorem Let a simple example introduce the first theorem. Suppose we want to predict the oil price for the following year by considering an estimator averages two independent expert opinions (expert A and expert B). If A predicts 50 € for a barrel and B predicts 60 €, and if the actual barrel price is 70 €, the average absolute error of the experts is 15 €, that is, $(20\text{ €} + 10\text{ €})/2$. By comparison, the estimator based on the mean opinion ($\bar{x} = 55$ €, denoted as the *crowd error*) also leads to a 15 € error. However, if the actual price lies between the experts' judgments (e.g., 58 € instead of 70 €), the experts' average error (5 €) becomes higher than the crowd error (3 €).

The crowd error is always equal or smaller than the average error for any convex loss function [Larrick and Soll (2006)]. Sociologists usually measure errors with absolute deviations (as in the above example) while mathematicians

[2] By contrast with a sociological or psychological effect

and engineers rather use squared deviations. Throughout the rest of this thesis, we will express the errors with square loss functions. When denoting $\bar{x}$ the mean of n opinions, and T the corresponding true answer, the errors' inequality is a mathematical result that can be expressed as :

$$\begin{aligned} \text{Crowd error} &\leq \text{Average error} \\ (\bar{x} - T)^2 &\leq \frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 \end{aligned}$$

where the left and right terms correspond respectively to the crowd error and the average error. Hong and Page (2008) suggested a first formulation of the wisdom of the crowd by interpreting the difference between the crowd error and the average error as the predictive diversity :

$$\begin{aligned} \text{Crowd error} &= \text{Average error} - \text{Predictive diversity} \\ (\bar{x} - T)^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - T)^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned} \tag{2.1-DIV}$$

In the above example, the predictive diversity of the experts is 25. When the actual price of a barrel is 58 €, the experts' average error is 34, leading to a crowd error of $34 - 25 = 9$. Hong and Page justify this equality by stating that « the wisdom of the crowd is equal parts ability and diversity». It should be mentioned that the equation (2.1-DIV) is nothing more than the variance decomposition in an ANOVA analysis[3] [Walker (2016)]. The *Diversity Prediction Theorem* assesses the crowd error for a particular instance, that is, knowing the average error and the predictive diversity of a particular sample. By contrast, the following theorem expresses the expected crowd error when information is available concerning the population bias and variance.

The Bias-Variance-Covariance Decomposition We now discuss the ways in which quantitatively the concepts of bias, variance and covariance of a population could affect the expected crowd error. To understand this, these three notions needs to be introduced in the context of opinion aggregation.

[3]The decomposition results more fundamentally from the Pythagorean theorem generalized to a Euclidean space.

The opinions x_i of an individual i can be systematically lower or higher than the ground truth T. That is, their average opinion (noted μ_i) is either above ($\mu_i > T$) or below ($\mu_i < T$) the true value. This systematic error is referred to as the *bias* : $b_i = \mu_i - T$. Individual opinions are also varying around the mean value, generating an additional noise often called the idiosyncratic noise. This noise is measured by the variance $v_i = E(x_i - \mu_i)^2$. Finally, opinions between pairs of individuals are not always independent and can vary in the same or opposite direction. This is measured by the opinion covariance between the individuals i and j : $C_{ij} = E[x_i - \mu_i][x_j - \mu_j]$.

In the case of n expressed opinions with an average bias $\bar{b}$, an average variance $\bar{V}$ and an average covariance $\bar{C}$, Hong and Page (2008) show that the expected value of the crowd error can be decomposed as a sum of these three components:

$$\begin{array}{rcccccc} \text{Expected Crowd error} & = & \text{Bias} & + & \text{Variance} & + & \text{Covariance} \\ E[(\bar{x} - T)^2] & = & \bar{b} & + & \dfrac{1}{n}\bar{V} & + & \dfrac{n-1}{n}\bar{C}. \end{array} \quad \text{(2.2-BVC)}$$

In particular, if we assume unbiased ($\bar{b} = 0$) and independent ($\bar{C} = 0$) opinions, the collective error goes towards 0 when the crowd becomes large. In other words, averaging a huge number of independent opinions from an unbiased opinion distribution will cancel out the idiosyncratic noise, and the mean opinion of the population will accordingly lie close to the ground truth.

Relaxing the assumptions The assumptions concerning the absence of bias and the independence between judgments are often very restrictive, in addition to the necessity of having a large group size. Therefore, academics started to investigate the effect of relaxing those assumptions on the crowd error. Ariely, Tung Au, Bender et al. (2000) showed that even small groups (from 2 to 6 judges in their study) could lead to substantial improvements when averaging opinions. Davis-Stober, Budescu, Dana et al. (2014) indicated that biased judgments generate a collective opinion that is generally far more accurate than the one of a randomly sampled member. Finally, the work of Lorenz, Rauhut, Schweitzer et al. (2011)—amongst others—questionned the independence assumption by investigating the effect of human influence on the crowd error. They carried out social experiments where the members are interconnected and share information about their judgment. By observing the crowd error, they discovered

that mutual influence can cause the collective to become highly inacurrate.

2.3 Analyzing social interconnections

We have seen that a sufficient condition to observe the wisdom of crowd effect is to aggregate independent and unbiased opinions. In practice, opinions are far from independent, due to the people's daily interactions with the others. Human beings are often part of an underlying offline or online network that dynamically alters their opinions over time. In particular, social networks are usually structured by a group of interconnected users or members. In this section, we explore the representation and analysis of networks by providing a brief introduction to graph theory and social network analysis.

2.3.1 Graph theory

Graph theory provides a wide variety of tools for extracting interesting properties from various kinds of networks. Mathematically, a graph $G(V, E)$ of order n is a structure that combines a set of n vertices $V = \{1, ..., n\}$ with a set of connecting *edges* $E = \{(i, j) | \, i, j \in V\}$. Figure 2.1 displays two graph instances of order 4. The left graph (a) depicts an undirected graph composed only of symmetrical relations. By contrast, the right graph (b) is made of directed *edges*. The latter may be used for modeling potential non-reciprocal relations. Graphs with directed edges are called digraphs or directed graphs.

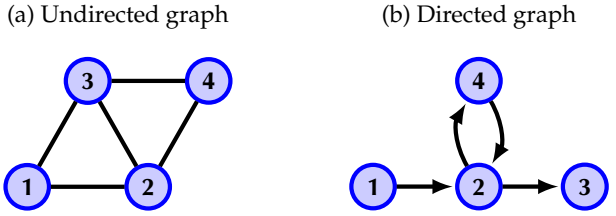


Fig. 2.1 Two examples of graphs with 4 nodes.

Some agents are more influential or popular in their community than others. One proxy of popularity often used is the *degree* of nodes in a graph. The *degree* of a particular node is defined as its number of incident edges. For directed graphs, a distinction is made between in-degree (incoming *edges*) and out-degree (outcoming *edges*).

They are many ways to connect n users when experimenting on the topic of social influence. To connect the users to the whole network, one has to

create for each of them a connection to the $(n-1)$ other agents. The resulting graph—labeled K_n—is called *complete graph* and is made of $n\,(n-1)/2$ edges. Figure 2.2 displays the complete graphs for groups going from 1 to 5 agents.

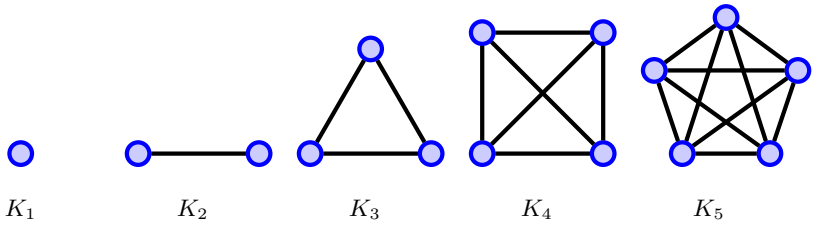


Fig. 2.2 The first five complete graphs.

Representation of graphs Nowadays, social networks are made of millions of nodes. The matrix representation of a graph is a common way to analyze or store various kinds of networks. In this thesis, we choose to work with *adjacency* matrices. They are square matrices of order n representing the existing connections between the n vertices. The matrix entries a_{ij} are set to 1 when an edge from vertex i to j exists, and to 0 otherwise. The adjacency matrices for the two examples of Figure 2.1 are given by :

$$A_{und} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad A_{dir} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

where A_{und} is the adjacency matrix of the undirected graph (a) and A_{dir} is the representation of the digraph (b). We observe that undirected graphs are generating symmetrical matrices, which is not necessarily the case for directed networks. Note that the number of nonzero entries reflects exactly the number of edges for a digraph and counts twice the edges present in an undirected graph.

Walks and paths We will see in Chapter 4 that influence networks are determined by the connectivity of the underlying graph. The connectivity of a graph is described by moves between connected nodes. This is captured by the notion of *walk* in a graph, which is defined as an alternating sequence of nodes and edges $(v_0\,, e_0\,, \ldots, v_k\,, e_k)$ where each edge e_i is incident to the surrounding nodes. For directed graphs, one requires in addition that all edges e_i composing the walk have the same orientation. Walks can contain two or more instances of the same node ($v_i = v_j$ for some $i \neq j$). Walks without duplicated

vertices are referred to as *paths*. The length of a path is defined as the number of edges composing the path, as shown in Figure 2.3. Besides, paths in which the direction of the edges are ignored are referred to as *semi-paths*. Two nodes are *connected* if there is a path between the two nodes in both directions.

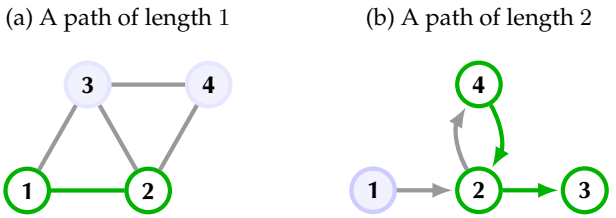


Fig. 2.3 Paths in undirected and directed graphs. Both paths are displayed in green.

2.3.2 Social Network Analysis

By harnessing methods and algorithm issued from graph theory, sociologists analyze complex structures of social networks to reach a better understanding of human behavior. In the field of *Social Network Analysis* (SNA), we define a group or network as a set of two or more individuals ($n \geq 2$) who share connections through social relationships. They can range from a small number of agents (e.g., groups of co-authors on a project) to very large collectives (e.g., entire research communities). Groups are often classified according to their shared similarities [Forsyth (2009)]. In this sense, *primary groups* are long-term groups characterized by face-to-face interactions. Typically, they bring together friends and family members. By contrast, *secondary groups* are shorter in duration, less emotionally charged and they reflect the ways in which the social agents are involved in their social environment. They include, for instance, the professional-to-professional network.

A social group or network can be rendered in the form of a graph where the nodes represent human agents and where the edges capture a particular relationship. The nature of the social connections varies from one network to another. As an example, the online network *Facebook* is mostly composed of symmetrical friendship connections. In the resulting *Facebook* graph, we observe the existence of an edge only when the two related nodes agree with the friendship relation. By contrast, the *Twitter* platform is composed of asymmetrical relations that distinguish the followers from the followees. As a result, undirected graphs are usually chosen to represent symmetrical relations while directed graphs (digraphs) are more suitable when the orientation of the con-

nection matters. It is worth noting that interconnections may also be more implicit. This is the case, for example, when agents share similar contents or when they comment on similar topics.

Psychologists and sociologists often make the distinction between small and large groups. The term *small* often refers to groups made up from 2 to 20 individuals [Shaw (1971)]. Their small size makes them tractable for a qualitative analysis and their underlying structure does not require advanced visualization tools. By contrast, larger concentrations of humans—which are sometimes represented by graphs featuring billions of nodes and edges—are untractable without designing proper measurements tools. The smallest groups consist in two members and are called *dyads*. Similarly, groups of three or four people are respectively denoted by *triads* and *tetrads*.

Sociograms The first quantitative methods to study social relationships were developed by the psychosociologist Jacob Levy Moreno, around 1932, at the New York State Training School for Girls in Hudson. The field that he coined *sociometry* analyzes the affective relationships appearing inside a group by the use of *sociograms*. A sociogram is a particular type of graph where the edges represent relations or preferences that agents express amongst themselves. Building a sociogram requires to determine a positive or negative criterion which depends on the context of the study. The agents are then asked to answer one or multiple questions related to the criterion. Two examples of questions might be: « Who do you think is the best leader in your network? » or « Whom in the group would you choose to keep a confidence? ».

Amongst others, Moreno considered the use of sociograms to assign residents to residential cottages and for educational purposes. In the latter case, he realized an experiment on pupils where he observed different clusters between the kids by asking about their seating preferences. Later, sociograms became a reference tool to analyze social relations inside working teams. In their work, Cross, Borgatti, and Parker (2002) show how to make good use of sociograms to facilitate the collaboration in top leadership networks. As another example, we present in Figure 2.4 (a) the hierarchical structure of an audit firm analyzed in the research work of Krackhardt (1996). Figure 2.4 (b) displays the associated sociogram where the directed edges represent affirmative reactions to the following question :

> « Who among the people here in the auditing group do you typically go to for help or advice when you encounter a problem or have a question at work? » [Krackhardt (1996)]

The sociogram highlighted one particular node (Nancy) with a high number of incoming edges, which allowed the manager to detect the unbalanced workload and to considerably improve the performance of the firm.

Usually, sociograms are used to identify nodes that present particular properties. Nodes with a high in-degree (e.g., Nancy in the previous case) are referred to as *stars* by the sociologists. They often correspond to leaders or popular members in a group. By contrast, agents who do not count any relationship amongst the network are characterized by the absence of incoming edges in the graph. These are represented by vertices called *isolated nodes*. Some nodes are very crucial to the connectivity of the network : these are called the *cutpoints*. When removing a cutpoint from the network, we directly disconnect multiple other nodes from the network. This is the case of Charles in Figure 2.4 (b).

Metrics When the group size becomes large, it is often difficult to draw a clear and readable sociogram, and sociologists have to consider more complex measurement tools. Table 2.1 presents three common properties used in graph theory that are particularly meaningful in social network analysis. *Triadic disclosure* is a determinent measure in the analysis of information diffusion and community organization [Granovetter (1973)]. It was first introduced by the sociologist Simmel (1908), who stated that, when two connections exist in a triad, there is a high probability of observing a third connection. This is commonly measured in graph theory by the clustering coefficient of a graph :

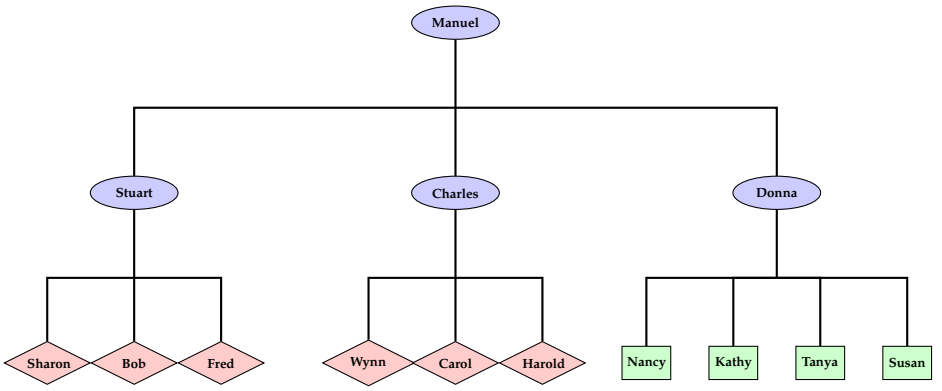
$$\text{clustering coefficient} = \frac{\#\text{ closed triads}}{\#\text{ connected triads}}$$

where a closed triad refers to a complete graph K_3 and where $\#$ designates the cardinality of a set.

Graph Theory	Social Network Concept
Clustering coefficient	Triadic closure
Centrality coefficient	Importance, Popularity
Assortativity coefficient	Homophily

Table 2.1 Graph theory coefficients conventially used in social network analysis.

(a) Organizational chart



(b) Sociogram

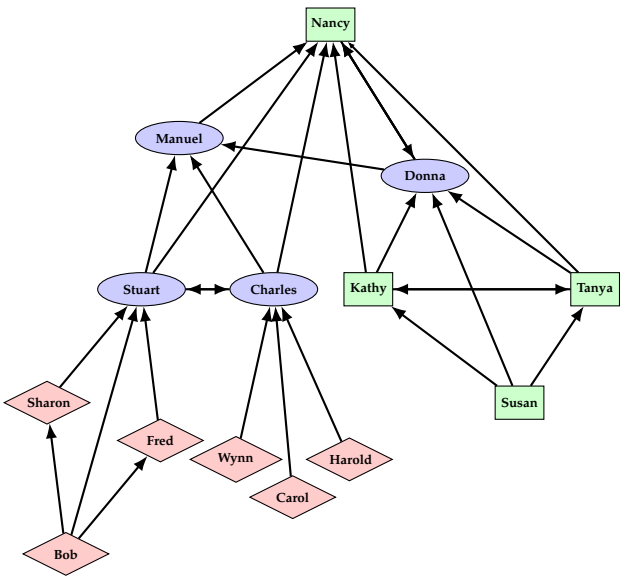


Fig. 2.4 Organizational chart and the associated sociogram from an auditing team [Krackhardt (1996)]. Managers are represented by ellipses, staff auditors by diamonds and secretaries by rectangular boxes. A line from A to B indicates that A seeks advice from B.

Influential and popular actors are detected by analyzing the *centrality* of the nodes. There are different ways of defining the centrality of a node. The stars presented in the field of sociometry are nodes with a high *centrality degree*, measured by the degree of each node. By contrast, the *closeness centrality* [Bavelas (1950)] assesses the centrality of a node in terms of average *distances*:

$$\text{closeness centrality coefficient} = \frac{1}{\text{average distance to the other nodes}}$$

where the distance between two nodes is defined as the length of the shortest path connecting them.

Finally, the assortativity coefficient of a network measures in some way the similarity between connected agents. The assortativity coefficient as suggested by Newman (2002) is defined as the Pearson correlation coefficient of the degrees between pairs of connected vertices. In sociology, the fact that similar individuals tend to connect to each other is designated by the notion of *homophily*. As an example, the assortativity as described above measures whether high degree agents have a preference for creating links with other high degree nodes (and conversely).

2.4 Conducting behavioral research on crowdsourcing platforms

Analyzing opinions of connected agents can be carried out either by observing the people's behavior on existing social networks (*in vivo studies*, i.e., under natural conditions) or by developing rigorous survey methodologies and laboratory studies (*in vitro studies*, i.e., under laboratory conditions). While the former are often costless for the researcher, the latter requires high monetary costs and time delays to collect a large enough sample [Kittur, Chi, and Suh (2008)]. The use of *crowdsourcing platforms* is one cost-effective way to conduct psychosocial experiments and allows to gather many users for a limited period of time. These platforms usually offer tasks to be completed by the online crowd for a small fee. In 2005, Jeff Howe and Mark Robinson (two editors at *Wired Magazine*) were the first ones to write about *crowdsourcing*. This neologism designates the act of outsourcing tasks to the crowd. The existing crowdsourcing websites provide a way to expand from the small scale of laboratory experiments to the large scale of the online world. Unlike in the lab, once the experiment is prepared, data gathering is automatic. Platforms gather two types of actors : *requesters* and *workers*, who can be seen respectively as employers and employees. Two popular platforms are the *Amazon's Mechanical*

Turk[4], especially designed for simple tasks, and *Crowdflower*[5], which dispatches the tasks on multiple channels with advanced quality controls.

While one could think that computer-mediated social interactions do not reflect real-life interactions, more and more social interactions take place online. It thus makes sense to focus on these types of interaction for their own sake. Moreover, online experiments give the opportunity to easily reach beyond the WEIRD (Western, Educated, Industrialized, Rich, And Democratic) sample of the human population. Henrich, Heine, and Norenzayan (2010) show in their work why the WEIRD sample, which is customarily used in social psychology, is not representative of the human population. Conducting behavioral research on crowdsourcing platforms, however, requires great care in terms of experimental design. Mason and Suri (2012) offer interesting insights with regard to the way of properly considering the workers' contributions for behavioral studies. As an example, they suggest interesting ways to guarantee minimum data quality, such as duplicating tasks between multiple workers, including tasks with verifiable answers, and filtering very low entropy patterns. In the study of opinion dynamics, researchers are looking for ways to create interactions between workers. There is therefore a need for synchronous experimental conditions. The framework suggested by Mason and Suri (2012) allows to launch synchronous studies by following three key courses of action : building panels, notifying workers and conceiving a waiting room. A panel consists in the record of past workers who are interested in participating in future experiments. The notification send to the workers should include the times at which a new experiment is taking place. Finally, the waiting room helps to deal with the differences in the times at which the workers connect themselves on the platform. The two next chapters explore theories and models based on behaviors that were observed during synchronous crowdsourcing experiments.

[4] https://www.mturk.com/, last visit : 17th July 2017
[5] https://www.crowdflower.com/, last visit : 17th July 2017

Chapter 3

The Unpredictability of Human Judgment Revision

The two following chapters—Chapter 3 and Chapter 4—aim to investigate the prediction efficiency of theoritical opinion dynamic models when they are tested against a real dataset. The present chapter acts as a preliminary chapter to Chapter 4 as it highlights the necessity of estimating the unpredictable variations that occur in human judgment revision. Estimating the level of *unpredictability* provides a way to derive a lower bound on the prediction errors caused by classical judgment revision models. We will first present an experimental framework which, based on simple games, is designed to reveal the variations found in human opinion dynamics. We will then provide an estimation of this *intrinsic noise* and show how it may impact the errors made by opinion revision models.

3.1 Opinion formation : a stochastic process

Trying to describe how individuals revise their judgment when subject to social influence has a long history in the psychological and social sciences. These works were originally developed to better understand small group decision-making. This occurs for instance when a jury in civil trials has to decide the amount of compensation awarded to plaintiffs [Hastie, Penrod, and Pennington (1983); Horowitz, ForsterLee, and Brolly (1996)]. Various types of tasks have been explored by researchers. These include the forecasts of future events, e.g., predicting market sales based on previous prices and other cues [Fischer and Harvey (1999)], the price of products [Schrah, Dalal, and Sniezek (2006)], the probability of event occurrence [Budescu and Rantilla (2000)], such as the

number of future cattle deaths [Harvey and Fischer (1997)], or regional temperatures [Harries, Yaniv, and Harvey (2004)].

Some *in vivo* large scale online experiments were devoted to understand how information and behaviors spread in online social networks [Centola (2010)]. Others focused on determining which sociological attributes such as gender or age were involved in social influence processes [Aral and Walker (2012)]. Complementarily to the *in vivo* experiments, other recent studies used online *in vitro* experiments to identify the micro-level mechanisms susceptible to explain the way social influence impacts human decision-making [Chacoma and Zanette (2015); Mavrodiev, Tessone, and Schweitzer (2013); Moussaïd, Kämmer, Analytis et al. (2013)].

The predictability assessment is a necessary step to grow confidence in our understanding and in turn uses this mechanism as a building block to design efficient online social systems. Revising judgments after being exposed to others' judgments takes an important role in many online social systems such as recommendation system [Dellarocas (2003); Pope (2009)] or viral marketing campaign [Bessi, Coletto, Davidescu et al. (2015)] among others. Human judgment is such a complex process that no model can take all its influencing factors into account. The prediction accuracy of a model is limited to the extent the opinion revision process is a deterministic process. However, there is theoretical and empirical evidence showing that the opinion individuals display is nondeterministic.
Various complex phenomena are taking place in the human system and affects inexorably the opinion formation process. Behavioral research showed that human subjects are influenced by most of the uncertainties present in their daily life. These observations directly highlight the stochastic nature of the nervous system [Ma, Beck, Latham et al. (2006)]. Sensory marketing experts showed that senses affect the perception of consumers and influence their judgments on products. Visual perception is known as the most complex pattern recognition process [Kersten and Yuille (2003)] and is often represented by Bayesian models. Time and memory also alter opinions. For instance, Vul and Pashler [Vul and Pashler (2008)] confirmed that when participants are asked to provide their opinion twice with some delay in between, they provide two different answers. Scholars clearly emphasize the importance of probabilistic models when retrieving information from memory [Steyvers, Griffiths, and Dennis (2006)].

It appears from these studies that the precision of opinion dynamic models is limited by the intrinsic variation in the human judgment process. We propose

in the following sections an experimental framework to analyze the degree of unpredictability in online communities. To the best of our knowledge, it is the first time that the unpredictable variations appearing in judgment revision processes are quantitatively assessed.

3.2 An online experimental framework

Our study aims at understanding the behavior of socially connected agents. Unlike most related studies, our research is web-based, where we observe interaction of participants on a self-made website specifically designed to study social influence. *In vitro* studies have the advantage of providing the micro-level data driving the collective dynamics. Another asset of *in vitro* approaches is the possibility to differentiate social influence from other confounding factors (such as homophily) thanks to the anonymity of influencing individuals [Aral, Muchnik, and Sundararajan (2009); Shalizi and Thomas (2011)]).

Collecting participants is achieved by linking our website to an existing online community. We obtain a cross-section of members via an external crowdsourcing platform. Among the available platforms, *Crowdflower* is considered as one of the most efficient crowdsourcing website for research, characterized for its efficient quality checking and the presence of a massive number of worldwide contributors. Currently, the Crowdflower community is formed by more than 5 million workers[1] from around the world. There is a need to encourage the participants to behave at their level best. This requires a financial incentive process adjustable to the performance of the contributors. Crowdflower distributes their tasks to different partners channels. Therefore, the crowdsourcing platform includes people of many ages and various countries. Most of the on-demand workers declare to be from USA (18%), India (12%) and UK (12%).

3.2.1 Designing tasks in uncontrolled environments

Building crowdsourcing jobs to analyze online interactions is not a straightforward process. Our study aims to design tasks that reveal how opinions evolve as a result of social influence. In contrast to a face-to-face experiment, building social tasks on crowdsourcing platforms entailed dealing with a new set of issues. Online design is more restrictive in the sense that the players are not in a controlled environment.

[1] Number published on the Crowdflower website by Renette Youssef, (2014, September 14). Crowdflower blog. Retrieved from https://www.crowdflower.com

Suitable tasks have to satisfy several constraints. First, to finely quantify influence, the tasks ought to allow the evolution of opinion to be *gradual*. Numbers are chosen as the way for participants to communicate their opinion. Multiple choice questions with a list of unordered items (e.g., choosing among a list of holiday locations) are discarded. Along the same lines, the evolution of opinion requires *uncertainty* and *diversity* of a sufficient magnitude in the initial judgments. The tasks are chosen to be difficult enough to obtain this diversity. Thirdly, to encourage serious behaviors, the participants are rewarded based on their success in a task, that therefore takes the form of an online game. This required the accuracy of a player to be computable. Games are selected to have an ideal opinion or *ground truth* which served as a reference. We reject therefore subjective questions involving for instance political or religious opinions. Additionally, the task has to satisfy two other constraints related to the online context where, unlike face-to-face experiments, the researcher cannot control behavioral trustworthiness. Since the educational and cultural background of participants is a priori unknown, the game has to be *accessible*, i.e., any person who could read English has to be able to understand and complete the game. For instance, games could not involve high-level mathematical computations. Lastly, to anticipate the temptation to *cheat*, the solution to the games has to be absent from the Internet or any other external source. Therefore, we discard questions such as estimating the population of a country.

3.2.2 Games that generate unidimensionnal opinions

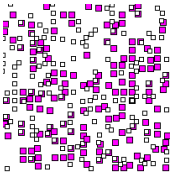
In the light of the above criteria, two types of games are designed. In the *gauging game* (Figure 3.1-a), the pictures are composed of 3 colors and participants estimate the percentage as a number between 0 and 100 of the same given color in the picture. In the *counting game* (Figure 3.1-b), the picture is composed of between 200 and 500 many small items, so that the player could not count the items one by one. The players then have to evaluate the total number of these items as a number between 0 and 500.

Participants start the experiment by joining groups of 6 participants. Their task is to successively play 30 games of the same sort related to 30 distinct pictures. The more the player's answers lie near to the truth, the more points they accumulate. At the end, the points are linearly converted to money and transferred to their *Crowdflower* account. These games lead to the creation of two datasets – *Dataset A* (detailed in the current section) and *Dataset B* (detailed in the next section).

(a) Example of a *counting game*

What percentage of the red color do you see in the image?

(b) Example of a *counting game*



How many items do you see in the image?

Fig. 3.1 Participants face either *gauging games* or *counting games*.

3.2.3 Connecting participants : Dataset A

The experiment is divided in several steps. During an instruction phase, players who would like to participate are asked to be familiar with the experiment process and are invited to read carefully the game rules, retracing inter alia how many points could be earned at each round. Subsequently, the experiment requires data about the players' social situation in order to evaluate possible biases. Not only do they have to notice a fluent English understanding, but they also need to fill out personal fields related to their age range and work experience. Thereafter, the participants are redirected to a personality webpage. In this respect, the aim is to provide a brief measure of the players' personality traits. At the end, they are brought to a waiting room and the playing phase can start shortly after. This is further detailed in Appendix A.

A game always consists of 3 rounds. The picture is kept the same for all 3 rounds. In each round, the participant has to make a judgment. During the first round, each of the 6 participants provide their judgment, independently of the other players. During the second round, each participant anonymously receives all other judgments from the first round and provides their judgment again. In other words, interactions can be represented by the complete graph K_6 where each player has access to the judgments of the whole network. The third round is a repetition of the second one. This process is illustrated in Figure 3.2 and brings us to the creation of the first dataset.

Dataset A - (*Groups of 6 real players*) The data were collected during July, September and October 2014. Overall, 654 distinct participants took part in the study (310 in the *gauging game* only, 308 in the *counting game* only and 36 in both). In total, 64 groups of 6 players completed a *gauging game*, while 71 groups of 6 participants completed a *counting game*. According to their IP addresses, participants came from 70 distinct countries. Participants mostly

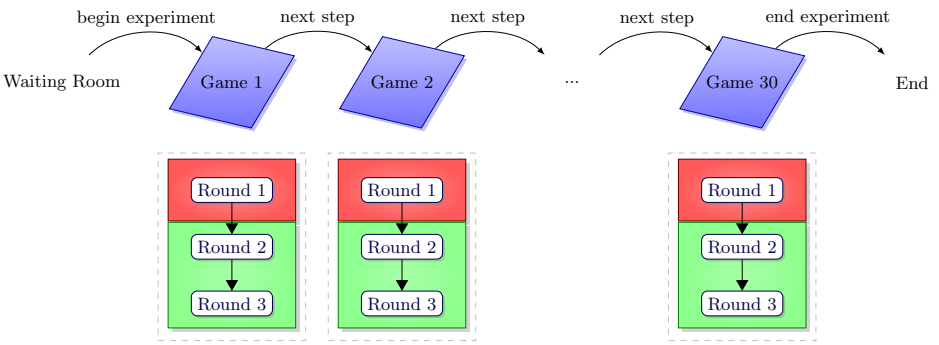


Fig. 3.2 Global overview of the crowdsourcing experiments. The first rounds (in red) are lone rounds. The second and third rounds (in green) are social rounds.

originated from 3 continents : $1/3$ from Asia, $1/2$ from Europe and $1/6$ from South America.

Most players completed most of the 30 games and played trustworthfully. The others were ignored from the study via two systematic filters. First, the predictions reported in present study concern only the participants who completed more than 15 out of the 30 games. This ensures that the number of games used for later parameter estimations is homogeneous over all participants. The median number of fully completed games per person was 24 for the *gauging game* and 27 for the *counting game*. Lower numbers are possibly due to a loss of interest in the task or connexion issues. The first filter lead to keep 68% of the participants for the *gauging game* and 71% for the *counting game* (Figure 3.3). Secondly, the predictions were only made on judgments of trustworthy participants. Trustworthiness was computed via correlation between participant's judgments and true answers. Most players carried out the task truthworthfully with a median Pearson correlation of 0.85 for the *gauging game* and 0.70 median correlation for the *counting game*. A few people either played randomly or systematically entered the same aberrant judgment. A minimum Pearson correlation threshold of 0.61 for the *gauging game* and 0.24 for the *counting game* were determined using Iglewicz and Hoaglin method based on median absolute deviation [Iglewicz and Hoaglin (1993)]. The difference between the two thresholds is due to the higher difficulty of the *counting game* as expressed by the difference between median correlations. This lead to keep 91% and 96% of the participants that had passed the first filter (Figure 3.4).

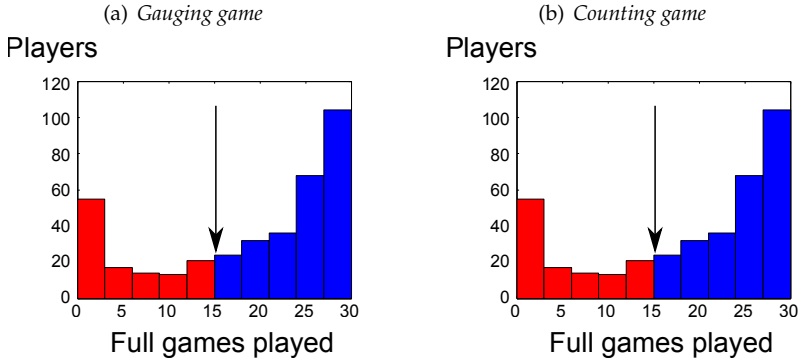


Fig. 3.3 Filters for players' participation. Histograms of the number of full games played by participants. Rejected players are displayed in red while kept players are in blue, to the right of the black arrow.

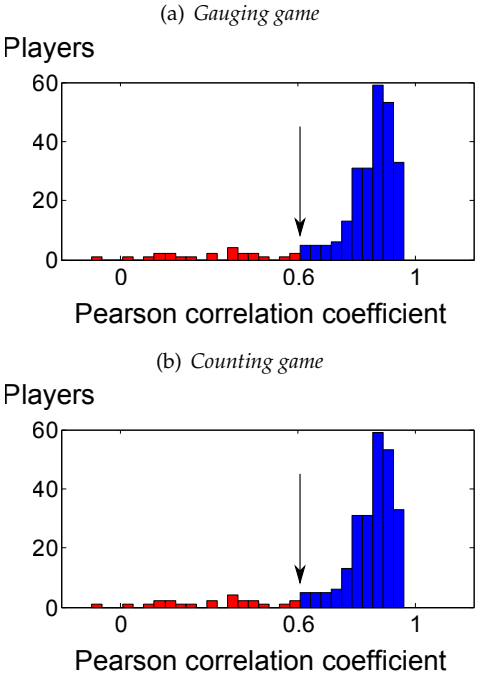


Fig. 3.4 Filters for participants' trustworthiness. Histograms of the correlations between judgments and true values for each participant. Rejected participants are displayed in red while kept participants are in blue, to the right of the black arrow.

It should be acknowledged that the selection procedure may have led to sample bias. This could be due to self-selection : some people choose to participate in the experiment and others do not. The fact that the study was carried out online is another factor that could bias the sample. The *a posteriori* filters may also be a source of bias. These are common issues in the behavioral sciences. Possibly, the nature of the study will have appealed to certain types of people and not others. Although that could have biased the characteristics of the sample, we are unaware of any empirical evidence suggesting that people who like participating in this kind of tasks are more or less susceptible to social influence.

3.3 Quantifying the degree of unpredictability

In this section, we want to measure the variation in the judgment revision process that would occur if a participant is exposed twice to identical pictures and in which the set of initial opinions happens to be identical. We propose a formal definition of these intrinsic variations and derives an expression to quantify the degree of unpredictability based on an adapted experiment process. For the sake of clarity, we first focus on the first two rounds of the games, before extending our analysis to the third round.

3.3.1 The intrinsic variation

When a participant takes part in a game, their second actual judgment depends on several factors : their own initial judgment $x_i(1)$, the vector of initial judgments from other players denoted as $x_{others}(1)$ and the displayed picture. As a consequence, we model the second round judgment of a participant as

$$x_i(2) = f_i^1(x_i(1), x_{others}(1), picture) + \eta, \tag{3.1-BEST2}$$

where f_i^1 describes how a participant revises their judgment on average depending on their initial judgment and external factors. If it was known, the function f_i^1 in equation (3.1-BEST2) would provide the best prediction[2] regarding judgment revision. By definition, no other model can be more precise. The term $picture$ corresponds to the influence of the picture on the final judgment. The quantity η captures the intrinsic variation made by a participant when making their second round judgment despite having made the same initial judgment and being exposed to the same set of judgments and identical picture.

Two assumptions are made on the random variable η. Firstly, the standard deviation of η is assumed to be the same for all participants in the same

[2]That is, for a prediction efficiency assessed by mean square error.

game, denoted as std(η). The standard deviation std(η) measures the root mean square error between f_i^1 and the actual judgment $x_i(2)$; this error measures by definition the intrinsic unpredictability of the judgment revision process. Secondly, two variables η and η' drawn from the random intrinsic variation for two replicated games are assumed to be independent, which results in $\mathbb{E}(\eta'\eta) = \mathbb{E}(\eta')\,\mathbb{E}(\eta)$.

It follows from the definition of the revision function f an additional property on the expected value of the intrinsic variation:

Proposition 1. *The random variable η has zero mean.*

Proof. Assume by contradiction that $\mu = \mathbb{E}(\eta) \neq 0$.
Denote $\tilde{x}_i(2) = f_i^1(x_i(1), x_{others}(1), picture)$. Then, we can show that $\tilde{x}_i(2) + \mu$ would be a better model for $x_i(2)$ than $\tilde{x}_i(2)$. Indeed, the expected square prediction error would be

$$\begin{aligned}
\mathbb{E}\left((x_i(2) - (\tilde{x}_i(2) + \mu))^2\right) &= \mathbb{E}\left((x_i(2) - \tilde{x}_i(2))^2\right) - 2\mu\mathbb{E}\left(x_i(2) - \tilde{x}_i(2)\right) + \mu^2 \\
&= \mathbb{E}\left((x_i(2) - \tilde{x}_i(2))^2\right) - \mu^2 \\
&< \mathbb{E}\left((x_i(2) - \tilde{x}_i(2))^2\right),
\end{aligned}$$

where we used the fact that $\eta = x_i(2) - \tilde{x}_i(2)$. This contradicts the definition of f_i^1 as the best model.

□

3.3.2 Assumptions on the revision process

The function f_i^1 that describes the revision process is unknown, but it is reasonable to make the two following assumptions.

Splitting Assumption First, the function f_i^1 is assumed to be a sum of (i) the influence of the initial judgments and (ii) the influence of the picture. According to a parameter $\lambda \in [0, 1]$, the f_i^1 splits linearly into two components :

$$\begin{aligned}
&f_i^1(x_i(1), x_{others}(1), picture) = \\
&\qquad \lambda\, g_i^1(x_i(1), x_{others}(1)) + (1 - \lambda)\, h_i^1(picture),
\end{aligned} \tag{3.2-SPLIT}$$

where g_i^1 represents the dependence of the second round judgment on past judgments while h_i^1 contains the dependency regarding the picture. The parameter λ weights the relative importance of the first term compared to the

second and is considered unique for each particular type of game. When $\lambda = 1$, the effect of the task is ignored in the revision process. By contrast, higher values of λ corresponds to contexts where the task influence prevails over the group influence. Introducing this additional parameter in our model makes the estimation process more conservative than if one ignores the picture's effect. We acknowledge that alternative approaches may be considered to incorporate the task influence, including, among others, a non-linear decomposition or a dependence on the picture's complexity.

Shifting Assumption It is further assumed that if the initial judgment $x_i(1)$ and the others' judgment at round 1 are shifted by a constant shift, the g_i^1 component in the second round judgment will on average be shifted in the same way, in other words, it is possible to write

$$g_i^1(x_i(1) + s, x_{others}(1) + s) = g_i^1(x_i(1), x_{others}(1)) + s, \qquad \text{(3.3-SHIFT)}$$

where s is a constant shift applied to the judgments. We acknowledge that the shift invariance is a strong assumption and does not especially hold when the extreme opinions are lying close to the boundary. This boundary effect can, however, be alleviated by ignoring instances for which the lowest or highest opinion appears too close to the interval limits. This is implemented in a second crowdsourcing experiment described in the following section.

3.3.3 Back to the experimental conditions : Dataset B

We now slightly adapt the previous experiment to create duplicated initial conditions. In that sense, we design 20 games (among the 30 games) to form 10 pairs of replicated games with an identical picture used in both games of a pair. To make sure the participants do not notice the presence of replicates, the remaining 10 games are distributed between the replicates. This is illustrated at Figure 3.5. Games 1 to 10 are the duplicated ones and games A to J are the unique ones. In addition, the judgments of five out of the six participants are now synthetically designed. The only human participant in the group is not made aware of this, so they would act as if they are playing with human players. The 15 synthetic judgments (5 participants over 3 rounds) which appear in the first instance of a pair of replicates are constructed by copies of past judgment made by real participants (for instance, from *Dataset* A). Opinions of real participants in the replicated experimental conditions are stored in a second dataset.

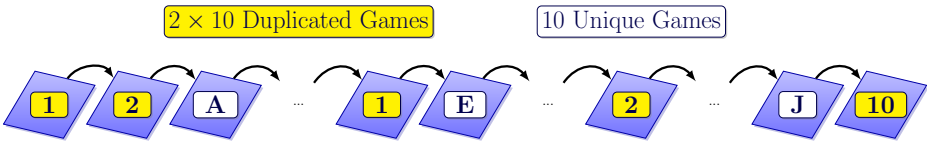


Fig. 3.5 Design of an experiment with duplicated pictures. Example of a 30-games sequence composed of ten duplicated pictures (1, 2, ... , 10) and ten unique pictures (A, B, ... , J).

Dataset B - (*Groups of 1 real player + 5 virtual players*) The data were collected during May and June 2015. Overall, 207 distinct participants took part in the study (113 in the *gauging game* only, 87 in the *counting game* only and 7 in both). The 120 *gauging game* players took part in 139 independent games while the 94 *counting game* players were involved in 99 independent games. Each independent game was completed by 5 synthetic participants to form groups of 6. The same filters as those used in *Dataset A* were applied to the participants. In addition, experiments containing opinions lying at less than 10% of the interval limits were automatically discarded. This led to keep 88% of the *counting games* end 80% of the *gauging games*.

Although we expose the real participant to replicated pictures, we cannot control their initial judgment. We handle this by shifting the 15 synthetic judgments in the second replicate. The shift made by the participant in the first round is synthetically applied to all other initial judgments. This maintains constant all the initial judgment distances in each replicate. Running example 1 illustrates how the shifting process generates the same set of initial judgments up to a constant in each pair of replicated games.

3.3.4 Deriving the unpredictability

Under assumptions (3.2-SPLIT) and (3.3-SHIFT), the adapted experimental framework provides a way of measuring the intrinsic variation (Proposition 2).

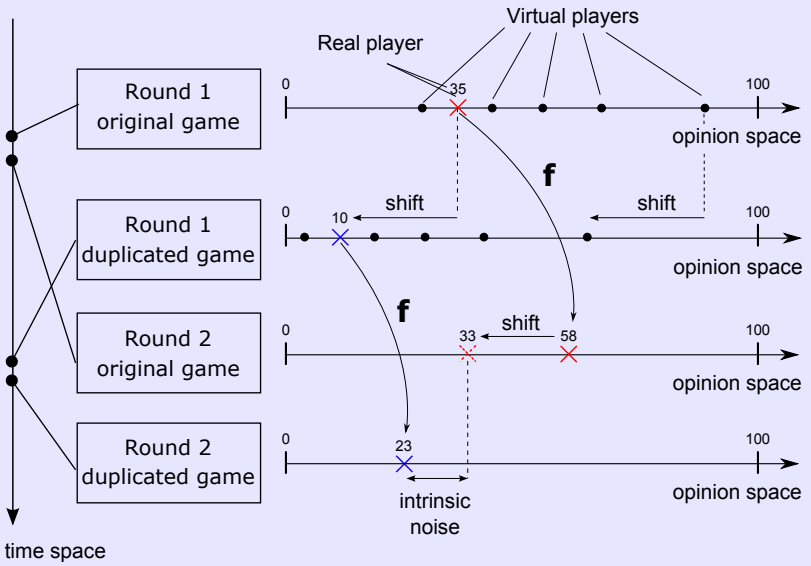
Proposition 2. *Under the splitting and shifting assumptions, the intrinsic variation estimation std(η) can be estimated by the root mean of*

$$\frac{1}{2}(x_i'(2) - x_i(2) - \lambda(x_i'(1) - x_i(1)))^2 \qquad \text{(3.4-UNP2)}$$

over all repeated games and all participants, where the prime notation is taken for judgments from the second replicated game in the experiment.

Running example 1 - The shifting process

To illustrate the shift applied to the synthetic judgment, we construct a running example based on opinions expressed in the range $0 - 100$:



One real player interacts with five virtual players. For a given picture, the initial opinion of the real player is 35. The player then revises their opinion to 58 after being exposed to the virtual players' opinions.

When exposed to the same picture later in the experiment, the same player is expressing an opinion of 10. In order to preserve the distances between initial judgments, a shift of $35 - 10 = 25$ is applied to the virtual players' opinions. The shift is computed in real time to match the variation of the real participant initial judgments between the two replicates.

We observe on the illustration a variation in judgment revision that can be computed as the distance between second round judgments (58 and 23), reduced by the shift in initial judgments. This variation is due to what we will call the *intrinsic noise*. The dotted cross (33) is not an expressed opinion and is displayed to show how the second round intrinsic variation will be computed using the initial shift. The same shift is also applied to all other second round judgments, and used to compute final round intrinsic variations.

Proof. The judgments made in two replicated games of a experiment by a same participants are described as

$$\begin{aligned} x_i(2) &= f_i^1(x_i(1), x_{others}(1), picture) + \eta, \\ x_i'(2) &= f_i^1(x_i'(1), x'_{others}(1), picture) + \eta', \end{aligned}$$

where the prime notation is taken for judgments from the second replicated game and η and η' are two independent draws of the random intrinsic variation. By design, the set of judgments are all shifted by the same constant :

$$\begin{aligned} x_i'(1) &= x_i(1) + s, \\ x'_{others}(1) &= x_{others}(1) + s, \end{aligned}$$

where $s = x_i'(1) - x_i(1)$ is known. According to the splitting assumption made on function f_i^1,

$$\begin{aligned} x_i(2) &= \lambda\, g_i^1(x_i(1), x_{others}(1)) + (1-\lambda)\, h_i^1(picture) + \eta, \\ x_i'(2) &= \lambda\, g_i^1(x_i'(1), x'_{others}(1)) + (1-\lambda)\, h_i^1(picture) + \eta', \end{aligned}$$

and the second round judgment made in the second replicate is then

$$x_i'(2) = \lambda \left(g_i^1(x_i(1), x_{others}(1)) + s \right) + (1-\lambda)\, h_i^1(picture) + \eta',$$

where the invariance by translation of g_i^1 was used. Taking the difference makes the unknown terms $g_i^1(x_i(1), x_{others}(1))$ and $h_i^1(picture)$ vanish to obtain

$$x_i'(2) - x_i(2) = \lambda s + \eta' - \eta. \tag{3.1}$$

Since η and η' have zero mean and are assumed to have equal variance, the theoretical variance of η is

$$\begin{aligned} \mathbb{E}(\eta^2) &= \frac{1}{2}\left(\mathbb{E}(\eta^2) + \mathbb{E}(\eta'^2)\right) \\ &= \frac{1}{2}(\mathbb{E}(\eta'^2) - 2\mathbb{E}(\eta'\eta) + \mathbb{E}(\eta^2) + 2\mathbb{E}(\eta'\eta)) \\ &= \frac{1}{2}(\mathbb{E}(\eta'^2 - 2\eta'\eta + \eta^2)) + \mathbb{E}(\eta'\eta) \\ &= \frac{1}{2}\mathbb{E}((\eta' - \eta)^2) + \mathbb{E}(\eta'\eta). \end{aligned} \tag{3.2}$$

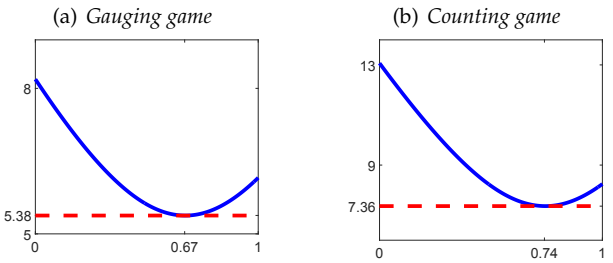


Fig. 3.6 Second round judgments : intrinsic variation as a function of the λ parameters, as given in equation (3.4-UNP2).

Moreover η and η' are assumed to be independent with zero mean, therefore, their covariance is null : $\mathbb{E}(\eta'\eta) = 0$. Consequently, $\mathbb{E}(\eta^2) = \frac{1}{2}\mathbb{E}((\eta' - \eta)^2)$ and using equation (3.1), the variance of η is empirically measured as the average of

$$\frac{1}{2}(x_i'(2) - x_i(2) - \lambda s)^2$$

over all repeated games and all participants, with the shift value of $s = x_i'(1) - x_i(1)$. □

Since η is assumed to have zero mean, the function f_i properly describing the actual judgment revision process is the one minimizing std(η). Correspondingly, the constant $\lambda \in [0, 1]$ is set so as to satisfy this minimization.

From the first and second round judgments recorded in *Dataset B*, we provide an estimation of the intrinsic variation std(η) as a function of the λ parameters (Figure 3.6). The optimal values are found to be 0.67 and 0.74 and correspond to intrinsic unpredictability estimations of 5.38 and 7.36 for the gauging game and the counting game, respectively. Note that the intrinsic variation for the *counting game* has been rescaled by a factor of 5 to be comparable to the *counting game* range of $0 - 100$. As we will see later on, these values can be used to assess the prediction efficiency of opinion dynamic models.

3.3.5 Relaxing the assumptions on the intrinsic variation

One of the assumptions made on the intrinsic noise is that η and η' have the same variance and are independent for each participant.
Since function f_i^1 is unknown, it is not possible to directly test these assumptions. However, since pairs of replicates in the experiment are related to the same picture, it is unlikely that the covariance between η and η' would be

negative. If the covariance was positive, the quantity given in equation (3.1) would become a lower bound on the unpredictability threshold, as shown through equation (3.2). Finally, if η and η' did not satisfy the assumption of equal variance, the quantity in equation (3.1) would still correspond to the average variance $\frac{1}{2}\left(\mathbb{E}(\eta^2) + \mathbb{E}(\eta'^2)\right)$ which also represents the average intrinsic unpredictability, as seen in equation (3.2).

3.3.6 Extension to the third round judgment

The procedure to estimate third round intrinsic unpredictability varies slightly from second round estimation procedure. For third round judgments, a function with the same input as f_i^1, depending only on the initial opinions and the picture, cannot properly describe the judgment revision of one participant independently of the other participant's behavior. In fact, the third round judgments depend on the second round opinions of others, which results from the revision process of other players. In other words, in a situation where a participant was to be faced successively with two groups of other participants, who by chance had given an identical set of initial opinions, the second round judgments of the two groups could vary due to distinct ways of revising their judgments.

Formally, the deterministic part of the third round opinion of a participant $x_i(3)$ is fully determined as a function of their initial opinion $x_i(1)$, the initial opinion of others $x_{others}(1)$, the second round opinion of others $x_{others}(2)$ and the picture. So, the third round opinion can be written as

$$x_i(3) = f_i^2(x_i(1), x_{others}(1), x_{others}(2), picture) + \bar{\eta}, \tag{3.5-BEST3}$$

where $\bar{\eta}$ is the estimation of the intrinsic variation occurring after three rounds under the same initial judgments, same picture and same second round judgments from others. A same reasoning as Proposition 1 allows to show that the prediction error $\bar{\eta}$ at round 3 also has zero mean.
Under the same assumptions on function f_i^2 as those made on f_i^1, and analogously to equation (3.4-UNP2), std$(\bar{\eta})$ is measured by the root mean of

$$\frac{1}{2}(x_i'(3) - x_i(3) - \lambda(x_i'(1) - x_i(1)))^2 \tag{3.6-UNP3}$$

over all repeated games and all participants.

As for the second round judgments, we analyze the variation of the intrinsic noise estimation as a function of the parameter λ for the third round judgments. Figure 3.7 displays similar results as Figure 3.6. For the *gauging games* (resp.

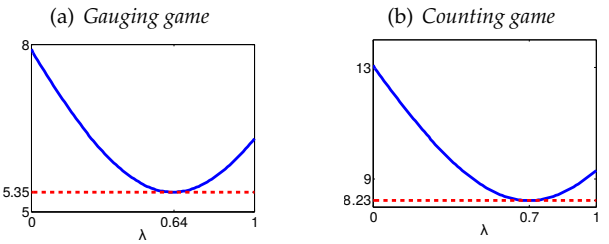


Fig. 3.7 Third round judgments: intrinsic variation as a function of the λ parameters, as given in equation (3.6-UNP3).

counting games), we observe that the intrinsic variation reaches a minimum of 5.35 (resp. 7.36) for a value of $\lambda = 0.7$ (resp. 8.23). We will see in the next section how these values can be seen as lower bounds for assessing model predictions.

3.4 From unpredictability to predictability

By conducting our research under replicated experimental conditions, we have shown how to estimate the intrinsic noise arising from simple online interactions. This last section aims at providing one useful application of this estimation in the context of prediction tasks related to opinion revision. To begin with, we suggest a baseline model that derives an upper bound for prediction errors generated by future opinion dynamic models. We then introduce two ratios for analyzing the prediction efficiency of these revision models.

3.4.1 The null model

In the first dataset (Dataset A), groups of 6 real participants are formed. Most of the participants actually update their opinion after being visually exposed to the opinions of others. When no prior information is available about specific participants, the reason of this update is hard to infer. In this case, the judgment revision mechanism is assumed to be unknown, and predicting constant opinions over time is the only option.

We call this baseline approach the *null model*. Any other opinion dynamic model is deemed to perform better. The null model therefore acts as an upper bound on the prediction errors of future models. Formally, it generates respectively two sets of predictions

$$\begin{aligned} \tilde{x}_i^{\text{null}}(2) &= x_i(1) \\ \tilde{x}_i^{\text{null}}(3) &= x_i(1) \end{aligned}$$

for the second and third round judgments. Intuitively, the model assumes that users are not influenced by the opinion of others and that they persist in keeping their initial opinion.

3.4.2 A measure for prediction efficiency

Models representing opinion formation mechanisms are generally described by a dynamical process that takes into account the initial opinions of connected agents (see for instance Hegselmann and Krause (2002)). Let F^1 and F^2 be functions that predict the opinions at a distance of 1 or 2 rounds (i.e., that provide predictions for round 2 and 3). While the equations (3.1-BEST2) and (3.5-BEST3) characterize the theoretical functions f^1 and f^1 that perfectly describe the revision mechanism, the functions F^1 and F^2 approximate the influence process and can be seen as models for which we want to assess the predictability. Just like the theorical functions, the functions F^1 and F^2 take into account the set of initial opinions and are related to a specific picture.

Again, we will first perform the analysis on the second round predictions. The opinion dynamic model provides an estimation $\tilde{x}_i^F(2)$ of the second round judgments:

$$\tilde{x}_i^F(2) = F_i^1(x_i(1), x_{others}(1), picture).$$

By contrast with the null model, the predicted opinion $\tilde{x}_i^F(2)$ is now potentially different from the player's initial opinion. The closer the predictions $\tilde{x}_i^F(2)$ lie from the observed opinions $x_i(2)$, the more predictive the model appears. In this work, *Root Mean Square Errors* (RMSEs) are chosen as a measure of the distance between estimates and observations. We suggest two ratios for analyzing the prediction efficiency of opinion dynamic models :

$$r_2^{noise}(F) = \frac{\text{std}(\eta)}{\text{RMSE}_2^F} \qquad \text{and} \qquad r_2^{imp}(F) = \frac{\text{RMSE}_2^{\text{null}} - \text{RMSE}_2^F}{\text{RMSE}_2^{\text{null}} - \text{std}(\eta)}.$$

The first ratio $r_2^{noise}(F)$ aims at estimating the proportion of prediction errors (RMSE_2^F) at round 2 that is due to the intrinsic unpredictability. The improvement ratio $r_2^{imp}(F)$ measures the improvement made by function F when

compared to a model that predict constant opinions over time ($\text{RMSE}^{\text{null}}$). This last ratio takes the unpredictability level as a reference.

The theoretical model f performs better by definition than any other opinion dynamic model ($\text{std}(\eta) \leq \text{RMSE}_2^F$). In turn, we expect a greater precision from a well-designed F function than from a baseline approach that simply replicates the initial conditions ($\text{RMSE}_2^{\text{F}} \leq \text{RMSE}_2^{\text{null}}$). Therefore, both ratios are expected to lie in the interval $[0, 1]$.
The value $r_2^{noise}(F) = 0$ reflects the theoretical situation where the human revision process is completely deterministic ($\text{std}(\eta) = 0$). When the ratio tends towards 1, one may conclude that most of the errors made by the model F actually result from the stochastic nature of judgment revision.
Analogously, when $r_2^{imp}(F) = 1$, no other model is supposed to offer better prediction estimates than F. On the contrary, lower values for $r_2^{imp}(F)$ correspond to poor prediction performances[3]. We reach the value $r_2^{imp}(F) = 0$ when the current model (F) is no more predictive than a function that simply ignores the revision mechanism. Note that the denominator does not depend on the model considered. This discrepancy – that we call the *scope for improvement* – is displayed at Figure 3.8 for the second round judgments. The null model variation ($\text{RMSE}_2^{\text{null}}$) is computed from *Dataset A*.

3.4.3 Limitations for the third round judgment

Expanding the previous reasoning to the third round judgments is not straightforward. As a reminder, since the initial opinions does not suffice in describing third round opinions, the function f_i^1 has been modified to take the second round opinions of others $x_{others}(2)$ as an additional input.
However, we want to perform predictions based solely on the initial judgments. In other words, we want to assess an opinion dynamic model that provides estimates

$$\tilde{x}_i^F(3) \quad = \quad F_i^2(x_i(1), x_{others}(1), picture).$$

that does not include information on $x_{others}(2)$, the second round judgments. This additional piece of information, taken into account by function f_i^2, necessarily makes the description of third round judgment more precise. As a consequence, the intrinsic variation estimated at equation (3.6-UNP3) is an underestimation of the actual intrinsic variation included in the prediction error of the consensus model. To summarize, the description of judgment

[3]Except for stubborn players that are not influenced by the group.

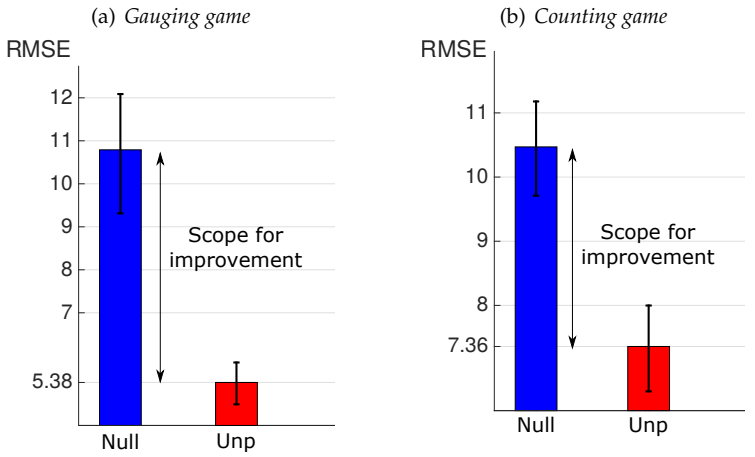


Fig. 3.8 The scope for improvement for the second round, located between the root mean square error (RMSE) of the null model and the intrinsic noise (Unp). For the *counting game*, values have been scaled by a factor of 5 to be comparable to the *gauging game*. For visualisation purposes, the y-values are displayed on different scale ranges between the two graphs.

revision does not strictly speaking provide the exact degree of intrinsic variation included in the final round prediction error, it is rather an underestimation.

Nonetheless, the intrinsic variation can still be used as a lower bound and we define a similar quality measurement for the third round estimations:

$$r_3^{noise}(F) = \frac{\text{std}(\eta)}{\text{RMSE}_3^F} \qquad \text{and} \qquad r_3^{imp}(F) = \frac{\text{RMSE}_3^{\text{null}} - \text{RMSE}_3^F}{\text{RMSE}_3^{\text{null}} - \text{std}(\eta)}.$$

Note that the null model variation is now computed on the last round. We define the *scope of improvement* from the underestimated intrinsic noise $\text{std}(\bar{\eta})$, as displayed at Figure 3.9.

In the next chapter, we will see how specific revision models can be used to estimate future opinions. Capturing the level of unpredictability allows to determine which part of the estimation errors are due to the stochastic human behavior. In addition, it does give an actual bound on the maximal possible improvement that can be obtained with models of a certain class—that is, those that can be described as (3.2-SPLIT). The intrinsic noise quantifies the highest prediction accuracy one can expect.

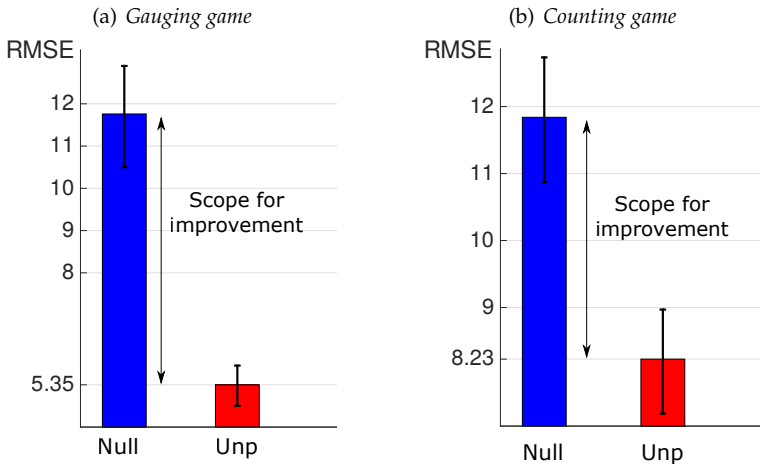


Fig. 3.9 The scope for improvement for the third round, located between the root mean square error (RMSE) of the null model and the intrinsic noise (Unp). For the *counting game*, values have been scaled by a factor of 5 to be comparable to the *gauging game*. For visualisation purposes, the y-values are displayed on different scale ranges between the two graphs.

3.5 Concluding remarks

The degree to which one may successfully intervene on a social system is directly linked to the degree of predictability of opinion revision. Because there must always be factors which fall out of the researcher's reach (changing moods or motivations of participants), part of the process cannot be predicted.

The present framework provides a way to assess the level of unpredictability of an opinion revision mechanism. This assessment is based on an experiment with hidden replicated tasks. Simple games are particularly suitable to study the revision mechanism. They especially allow the researcher to compute a ground truth related to each game. This is determinant to provide incentives for participants. We will see later that the availability of a ground truth allows to tackle interesting questions concerning the effect of influence on the group performance. Replicating identical experimental conditions is one of the major challenges when analyzing the degree of unpredictability. Our approach combines the staggering in time of duplicated tasks with simple assumptions on the revision mechanism. The proposed experiment type and validation method can in principle be generalized to any sort of continuous judgment revision.

Chapter 4

Analyzing Opinion Revision with Consensus models

The past two decades have witnessed a few attempts to confront simple models of opinion dynamics – the so-called *consensus models* – to real world data on collective human behaviors. These include conflicts and controversies on Wikipedia [Török, Iñiguez, Yasseri et al. (2013)] or how voters distribute their votes among candidates in elections [Bernardes, Stauffer, and Kertész (2002)]. However, the predictive power of these models was not assessed : the data used to calibrate the models also served to validate them. The easy access to large judgment revision database via online *in vitro* experiments opens up new opportunities for studying the predicting efficiency of consensus models.

4.1 Opinion dynamic models

In this first section, we discuss current mathematical formalizations that are usually employed to represent social phenomena. Opinion dynamic models have a long history in social science [French (1956)] and their behavior has been thoroughly analyzed in a theoretical way [Martin and Girard (2013); Olfati-Saber, Fax, and Murray (2007)]. The first contributions asked the question to know whether consensus could be reached amongst a group of agents. The underlying models are often referred to as *consensus models* [Blondel, Hendrickx, and Tsitsiklis (2009)] and express how agents revise their opinions as a function of the current distribution of judgments in the group.

We have seen that opinions are presented either as discrete choices or real quantities. Discrete choice models (as those described by Dodds and Watts (2004)

and Granovetter (1978)) directly link the discrete actions of an individual to the actions in their neighborhood. This allows to describe cascade propagation of behaviors. Presumably, before someone changes their actions, their opinion has to change as a result of the social stimuli. The second kind of models describe directly the opinions in a continuous space.
In this following paragraphs, we will once again consider the case of unidimensional and continuous opinions[1], i.e., $x_i \in \mathbb{R}$. We explicit the notion of *advice taking weights* and then introduce classical approaches for modeling influence and briefly discuss the issue of consensus within the group.

4.1.1 Advice taking weights

Consensus models are built to represent the formation of opinions in a group of interacting agents. Mostly, we just assume a finite number n of agents with no further constraints on the group size. The main factor entering in judgment revision models is the weight which individuals bestow on the judgments of others. This quantity is known as the *advice taking weight* [Harvey and Fischer (1997); Sniezek, Schrah, and Dalal (2004)] or the *weight of advice* [Gino (2008); Yaniv (2004a,b)]. It is represented by a number taking the value 0 when the individual is not influenced at all and 1 when they entirely forget their own opinion to adopt the vision of other individuals in the group. Mathematically, we denote the weights a_{ij}, which expresses how the agent i is influenced by the opinion of the agent j.

It has been observed that in a vast majority of the cases, the final judgment falls between the initial one and the ones from the rest of the group. Said otherwise, the weights often lie between 0 and 1. The condition $a_{ij} \in [0, 1]$ has been shown to sensibly improve the accuracy of decisions [Yaniv and Milyavsky (2007)]. A 20% improvement has been found in an experiment when individuals considered the opinion of another person only [Yaniv (2004a)]. However, individuals do not weigh themselves and others equally. They rather overweight their own opinions [Bonaccio and Dalal (2006)]. This has been coined egocentric discounting [Yaniv and Kleinberger (2000)].

Many factors affect influence. These include the perceived expertise of the adviser [Harvey and Fischer (1997); Soll and Larrick (2009)] which may result from age, education, life experience [Feng and MacGeorge (2006)], the difficulty of the task [Gino (2008)], whether the individual feels powerful [See, Morrison, Rothman et al. (2011)] or angry [Gino and Schweitzer (2008)], the size of the

[1] Note that multidimensional variants of these models have been thoroughly analyzed in other follow-up researches [Fortunato, Latora, Pluchino et al. (2005); Lorenz (2008)]

group [Mannes (2009)], among others. A sensitivity analysis allows to determine which factors most affect advice taking [Azen and Budescu (2003)].

4.1.2 Time dependence

Consensus models describe the variation of opinions over time ($x = x(t)$). Two alternative frameworks that incorporate time dependence consider either *continuous* or *discrete* variations. *Continuous* revisions of opinion depend on a time scale ranging over the real axis. Conversely, *discrete* time steps consider opinion shifts in terms of rounds $t \in \{0, 1, 2, \ldots\}$.

Time-varying weights can be considered in a similar frame of mind. A first case involves an explicit dependence of the weight ($a_{ij} = a_{ij}(t)$) that agent i puts on the judgment of agent j. This happens for instance when a decrease of influenceability over time can be detected, which is called *opinion settling*. Advice taking weights are sometimes modeled with an implicit dependence $a_{ij} = a_{ij}(x(t))$, that is, a dependence on time-varying opinions. As an example, one agent may suddenly decide to ignore an opinion that becomes increasingly distinct from their own judgment. More generally, there is a dependence regarding both variables $a_{ij} = a_{ij}(t, x(t))$.

4.1.3 Popular models

From now on, we focus on *discrete* time variations, though similar deductions could be made in the case of the continuous time scale. As aforementioned, advice taking weights are supposed to lie in the interval $[0, 1]$. For clarity, scholars further assume that

$$a_{i1} + a_{i2} + \ldots + a_{in} = 1$$

which fixes the value of a_{ii}. Opinion dynamics aim at expressing future opinions according to past information. A general formulation of opinion evolution is given by

$$x(t+1) = A(t, x(t))\, x(t) \tag{4.1-GEN}$$

where $A(t, x(t)) = (a_{ij}(t, x(t)))$ is an n by n matrix containing the nonnegative time-varying weights. The matrix $A(t, x(t))$ is right stochastic (the rows are summing to 1) under the previous assumption.

Models derived from (4.1-GEN) are not linear. Dropping the state dependence leads to simpler models

$$x(t+1) = A(t)\,x(t) \tag{4.2-TV}$$

also called *time-varying* consensus models. Note that these models can be interpreted in terms of inhomogeneous Markov chains. *Time-varying* models can be further simplified when influencabilities are considered constant over time:

$$x(t+1) = A\,x(t). \tag{4.3-CLASS}$$

In this last configuration, the evolution of opinion is characterized by a single and fixed matrix A. These models are sometimes referred to as *classical* models [Hegselmann and Krause (2002)].

Particular cases of these models were thoroughly analyzed from a theoretical perspective. Amongst these, the *bounded confidence* and the *Friedkin and Johnsen* models are the ones that attracted the most attention.

Bounded confidence Models with bounded confidence [Deffuant, Neau, Amblard et al. (2000); Hegselmann and Krause (2002)] assume that the parameters $a_{ij}(x)$ depend on the distance between a given opinion (x_i) and another influencing opinion (x_j). One well-known instance is the model suggested by Krause and Stöckler (1997), in which all the agents are equally influenced by the other agents present in their neighborhood. In such a case, the neighborhood of agent i is formally defined by

$$N(i,x) = \{j \in [1,n] \;\big|\; |x_i - x_j| \leq e_i\}$$

and depends on a threshold parameter ϵ_i. Consequently, the term $|N(i,x)|$ counts the number of agents in the neighborhood of agent i, and the model of Krause and Stöckler can be formulated as :

$$a_{ij}(x) = \begin{cases} |N(i,x)|^{-1} & \text{if } j \in N(i,x) \\ 0 & \text{if } j \notin N(i,x) \end{cases} \tag{4.4-KS}$$

and is characterized by state-dependent weights. Unlike *classical* and *time-varying* models, *bounded confidence* models are usually non-linear.

Friedkin and Johnsen Alternatively, the model suggested by Friedkin and Johnsen (1990) assumes that individuals always remain influenced by their initial opinion (sometimes called *prejudice*) over time. We obtain the FJ model

by slightly modifying the *classical* model in such a way that future opinions still depend on the initial vector $x(0)$:

$$x(t+1) = GA\,x(t) + (I - G)\,x(0), \tag{4.5-FJ}$$

where G is a diagonal matrix that represents the agents' interaction susceptibilities $g_i \in [0, 1]$. Such a matrix contains low elements ($g_i \approx 0$) when the agents are particularly commited to their initial judgment, and high susceptibilities ($g_i \approx 1$) when the members are predisposed to interpersonal influence.

4.1.4 Consensus

We end this section by briefly detailing a typical opinion dynamics problem concerning the formation of a consensus amongst agents. The problem is motivated by the willingness of the agents, in a social group, to reach an agreement when discussing an issue.

Given the agents' initial opinions, we try to derive conditions that lead asymptotically to a common opinion. In other words, we try to find conditions under which a value $c \in \mathbb{R}$ exists such that:

$$\lim_{t\to\infty} x_i(t) = c \qquad \forall i \in \{1, 2, \dots, n\}\ .$$

Consensus in the context of *classical* models (4.3-CLASS) have been initially investigated by DeGroot (1974) and later by Berger (1981). The authors derived necessary and sufficient conditions by considering matrix A as the transition probability matrix in a Markov chain. These conditions were based on the irreducibility and aperiodicity of the underlying chain. From their work, one can derive simple and intuitive conditions on the matrix A. As an example, a consensus is reached if every agent is positively influenced by any other agent ($a_{ij} > 0$ for all $i \neq j$). In a less restrictive way, the group is always able to achieve a consensus when the matrix A is primitive[2]. This is illustrate in Running Example 2. Similar conditions allow to extend these results to more complex models [Chatterjee and Seneta (1977); Krause (2000)].

[2]Primitive matrices are nonnegative square matrices for which there exists a k such that all the entries of A^k are strictly positive

Running example 2 - Sufficient conditions for consensus

Let A_{pos} and A_{prim} be two matrices representing the advice taking weights such that:

$$A_{pos} = \begin{pmatrix} 0 & 0.1 & 0.9 \\ 0.3 & 0 & 0.7 \\ 0.4 & 0.6 & 0 \end{pmatrix} \qquad A_{prim} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.3 & 0 & 0.7 \\ 0 & 0.8 & 0.2 \end{pmatrix}.$$

The two aforementioned conditions suffice to determine that these two matrices lead to consensus. In the first matrix, every agent puts a (strictly) positive weight on the other agents. The second matrix is a primitive matrix for the reason that A^2 is positive.

4.2 Attraction to the average

By contrast to the theoretical work on opinion dynamics, we investigate in the following the prediction efficiency of a *time-varying* consensus model in order to anticipate the evolution of opinions observed in the crowdsourcing experiment (described in Chapter 3). To the best of our knowledge, it is the first time the predictive power of a quantitative model of opinion dynamics is tested against a real dataset. By contrast to previous research on the topic, the model is validated on data that did not serve to calibrate it. This avoids favoring more complex models over simpler ones and prevents overfitting.

Recent online *in vitro* studies [Chacoma and Zanette (2015); Moussaïd, Kämmer, Analytis et al. (2013)] have led to posit that linear models may be appropriate to describe the way online members revise their judgment when exposed to judgments of others. Our model (4.6-CMOD) assumes that when an individual sees a set of opinions, their opinion changes linearly in the distance between their opinion and the average of the group opinions. There is recent evidence[3] supporting this assumption [Mavrodiev, Tessone, and Schweitzer (2013)]. In mathematical terms, $x_i(t)$ denotes the opinion of individuals i at round t and its evolution is described as

$$x_i(t+1) \;=\; x_i(t) \;+\; \alpha_i(t) \cdot (\bar{x}(t) - x_i(t)) \qquad \text{(4.6-CMOD)}$$

where $t = 1, 2$ and where $\bar{x}(t)$ is the mean opinion of the group at round t. The parameter $\alpha_i(t)$ is called *influenceability* and expresses how agents are attracted towards the average opinion. In other words, we assume that agents

[3] A discussion on a possible non-linear behavior is provided in Section 4.6.2.

are equally influenced by any other member of the group. This assumption is further supported by the fact that participants appears unidentified on the website to the other players. The model assumes that the influenceability may vary over time. In the context of the crowdsourcing experiment, the model associates two parameters for each participant : $\alpha_i(1)$, the influenceability after first social influence and $\alpha_i(2)$, the influenceability after second social influence.

Influenceabilities are directly mapped to the advice taking weights. When we write model (4.6-CMOD) in its matrix form :

$$\begin{aligned} x(t+1) &= x(t) \ + \ diag(\alpha(t)) \ (J-I) \ x(t) \\ &= A(t)\, x(t) \end{aligned}$$

we observe that the matrix of advice taking weights takes the form $A(t) = [I + diag(\alpha(t)) \cdot (J - I)]$, where $J = (1/n)\, \mathbf{1}\, \mathbf{1}^T$.

4.3 Interpreting the influenceability

In the present section, we discuss the values of the influenceability parameters when our model is fitted on *Dataset A* (described in Section 3.2.3).

The distribution of couples $(\alpha_i(1), \alpha_i(2))$ are obtained by fitting model (4.6-CMOD) to the dataset via a mean square minimization. This procedure amounts to likelihood maximization (ML) when the errors between model predictions and actual judgments are normally distributed and have zero mean [Bishop et al. (2006)]. By writing $X_i^p(t)$ the recorded opinion of player i for picture p in *Dataset A*, and $\bar{X}^p(t)$ the average opinion of the corresponding group, the minimization leads to:

$$\alpha_i(t) = \frac{\sum_{\text{pictures}} (X_i^p(t+1) - X_i^p(t)) \cdot \left(\bar{X}^p(t) - X_i^p(t)\right)}{\sum_{\text{pictures}} \left(\bar{X}^p(t) - X_i^p(t)\right)^2}. \tag{4.1}$$

It is expected that both $\alpha_i(1)$ and $\alpha_i(2)$ lie in the interval $[0, 1]$. However, the *influenceabilities* $\alpha_i(t)$ are voluntary not constrained in the range $[0, 1]$, allowing either repulsive behavior ($a_i < 0$) or even overcompensation ($a_i > 1$). If $\alpha_i(t) = 0$, the participant does not take into account the others and their opinion remains constant. If $\alpha_i(t+1) = \alpha_i(t)$, the influenceability does not

change in time and the opinion $x_i(t)$ eventually converges to the mean opinion $\bar{x}$. Instead, if $\alpha_i(t+1)/\alpha_i(t) < 1$, influenceability starts positive but decays over time, which represents opinion settling. When $\alpha_i(1)$ is nonzero, the ratio $\alpha_i(2)/\alpha_i(1)$ represents the decaying rate of influenceability.

4.3.1 Egocentric discounting and influence contraction

The marginal distributions of $(\alpha_i(1), \alpha_i(2))$ are shown in Figure 4.1. As expected, most values fall within interval $[0, 1]$, meaning that the next judgment falls between one's initial judgment and the group judgment mean. Such a positive influenceability has been shown to improve judgment accuracy [Yaniv and Milyavsky (2007)]. Most individuals overweight their own opinion compared to the mean opinion to revise their judgment with $\alpha_i(t) < 0.5$. This fact is in accordance with the related literature on the subject [Bonaccio and Dalal (2006)] and is called *egocentric discounting*.

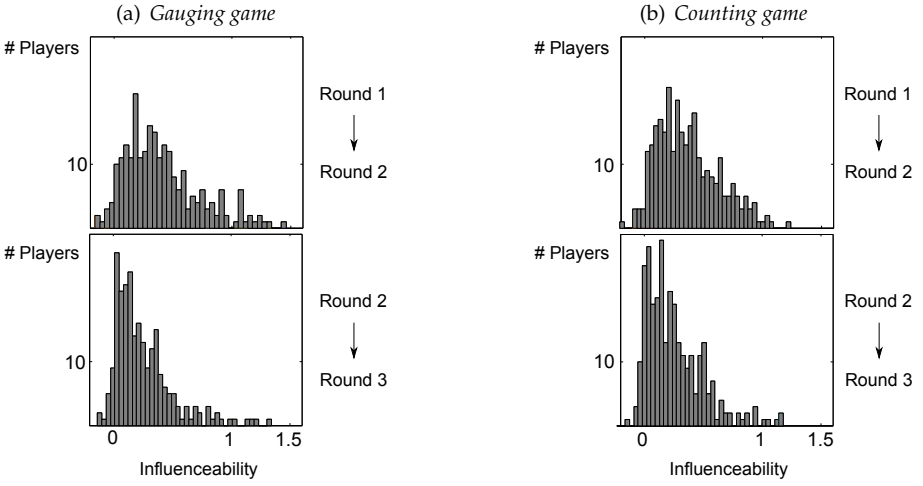


Fig. 4.1 Influenceability over rounds and games. Distributions of the $\alpha_i(1)$ influenceability after the first round and $\alpha_i(2)$ influenceability after the second round for the time-varying influenceability model (4.6-CMOD).

The plots in Figure 4.1 also display a small fraction of negative influenceabilities. One interpretation would be that the concerned participants recorded opinions very close to the average group opinion for multiple games. When this happens at one round, the opinion at the following round has a high probability to move away from the group average. This contributes to negative influenceabilities in

the participants' influenceabilities during the fitting process.

The distribution of influenceabilities of the population is subject to evolution over time. A two-sample Kolmogorov-Smirnov test rejects equality of distributions between $\alpha_i(1)$ and $\alpha_i(2)$ for both types of games (p-value $< 10^{-6}$, with KS distance of 0.25 and 0.23 for the *gauging game* and *counting game*, respectively). A contraction toward 0 occurs with a median going from 0.34 for $\alpha_i(1)$ to 0.18 for $\alpha_i(2)$ in the *counting game* and 0.32 to 0.20 in the *gauging game* (p-value $< 10^{-15}$ for a paired one-sided Wilcoxon sign-rank test). In other words, the participants continue to be influenced after round 2 but this influence is lightened. Figure 4.2 shows the discrepancy between the cumulated distribution functions over rounds.

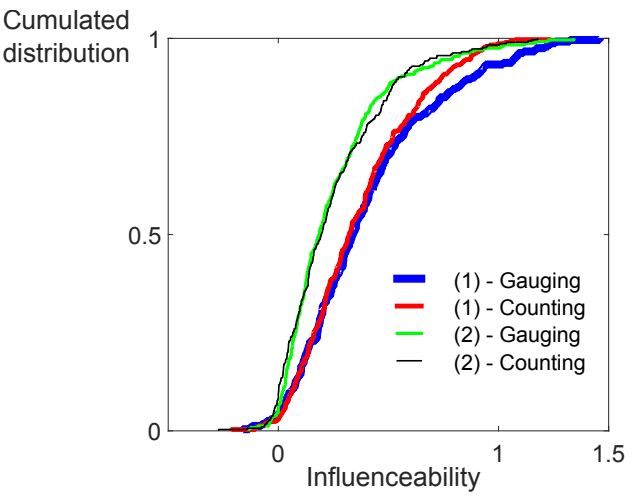


Fig. 4.2 Evolution of influenceabilities. Cumulated distributions of $\alpha_i(1)$ and $\alpha_i(2)$ for each type of games.

4.3.2 Influence and personality

An interesting research direction is to link the influenceability of an individual to their personality. To answer this question, we required the participants to provide information regarding their personality, gender, highest level of education, and whether they were a native English speaker. The questionnaire regarding personality comes from a research work by Gosling and Rentfrow [Gosling, Rentfrow, and Swann (2003)] and was used to estimate the five general personality traits. The questionnaire page is reported in Appendix A.2.

For each of the five traits, the participants rated how well they feel in adequacy with a set of synonyms (ratings $s \in \{1, \dots, 7\}$) and with a set of antonyms (rating $a \in \{1, \dots, 7\}$). This redundancy allows for testing the consistency of the answer of each participant. The participants who had a distance $(8 - a) - s$ too far away from 0 were discarded (threshold values were found using Iglewicz and Hoaglin method based on median absolute deviation [Iglewicz and Hoaglin (1993)]).

Partial Pearson's linear correlations are first reported between the individual traits measured by the questionnaire (see Figure 4.3). The correlation signs are found to be consistent with the related literature on the topic [Van der Linden, te Nijenhuis, and Bakker (2010)]. This indicates that our measure of the big five factors is trustworthy.

Partial correlations are then provided to link the personal traits to influenceability. As shown in Table 4.1, none of the measured personal traits is able to explain the variability in the influenceability parameter. The only exceptions concern gender and being native English speaker, with weak level of significance (p-value $\in [0.01, 0.05]$). However, these relations are consistent neither between types of tasks nor over rounds, so that they cannot be trusted. We conclude that the big five personality factors and the other measured individual traits have not proved to be relevant to explain the influenceability parameter. Finding appropriate individual traits to explain the influenceability remains an open question.

4.4 Predicting the opinion evolution

This section focuses in answering whether judgment revision models could be used to predict future decisions. Models are often validated on the data which served to calibrate the models themselves. This pitfall tends to favor more complex models over more simple ones and may result in overfitting. The model would then be unable to predict judgment revision from a new dataset.

4.4.1 Prediction scenarios

When one wishes to predict how a participant revises their opinion in a decision-making process, the level of prediction accuracy will highly depend on data availability. More prior knowledge on the participant should improve the predictions. When little prior information is available about the participant, the influenceability derived from it will be unreliable and may lead to poor

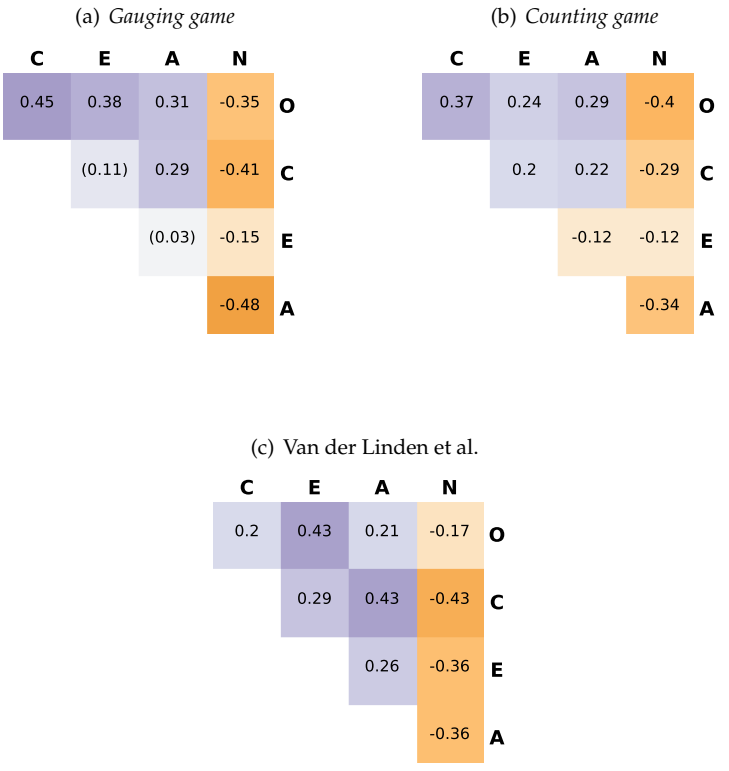


Fig. 4.3 Partial Pearson's linear correlations among the big five factors of personality. O: openness, C : calmness, E : extroversion, A : agreeableness, N : neuroticism. All p-values are significant (p-value < 0.05), except values inside brackets.

predictions. In this case, it may be more efficient to proceed to a classification procedure provided that data from other participants are available. These approaches are tested by computing the prediction accuracy in several situations reflecting data availability scenarios.

Scenario 0 - Baseline method In the worst case scenario, no data is available on the participant and the judgment revision mechanism is assumed to be unknown. In this case, predicting constant opinions over time is the only option. This corresponds to the null model against which the consensus model (4.6-CMOD) is compared.

Scenario 1 - Individual influenceability methods In a second scenario, prior data from the same participant is available. The consensus model can then be

	(a) Gauging game		*(b) Counting game*	
	$\boldsymbol{\alpha(1)}$	$\boldsymbol{\alpha(2)}$	$\boldsymbol{\alpha(1)}$	$\boldsymbol{\alpha(2)}$
O	-0.10	-0.04	0.07	-0.05
C	0.08	-0.06	-0.05	-0.09
E	-0.10	0.02	-0.01	-0.01
A	0.00	-0.02	0.10	0.07
N	-0.06	0.03	-0.10	0.01
Gen	-0.14*	-0.05	0.03	0.12*
Eng	-0.09	0.02	-0.03	-0.12*
Edu	0.05	-0.10	-0.04	-0.10

Table 4.1 Partial Pearson's linear correlations linking influenceability to the big five factors of personality. O: openness, C : calmness, E : extroversion, A : agreeableness, N : neuroticism, Gen: gender, Eng : native English speaker, Edu : highest level of education. Significance : *p-value < 0.05.

fitted to the data. Data ranging from 1 to 15 prior instances of the judgment process are respectively used to learn how the participant revises their opinion. Predictions are assessed in each of these cases to test how the predictions are impacted by the amount of prior data available.

Scenario 2 - Population influenceability methods In a final scenario, it is assumed that a large body of participants took part in a comparable judgment making process. These additional data are expected to reveal the most common behaviors in the population and enable to derive typical influenceabilities (population influenceabilities) by classification tools. For this last scenario, there are two possibilities. First, assume that prior information on the participant is available. In this case, the influenceability class of the participant is determined using this information. In the alternative case, no prior data is available on the targeted participant. It is then impossible to discriminate which influenceability class they belong to. Instead, the most typical influenceability is computed for the entire population and the participant is predicted to follow this most typical behavior.

4.4.2 The crossvalidation process

The goal of the crossvalidation procedure is to assess the prediction efficiency of model (4.6-CMOD) on a set that does not serve to calibrate the parameters. The set of first round judgments of this game are supposed to be known to initialize the model. This includes the initial judgment from the participant targeted for prediction and the initial judgments from the five other participants who possibly influenced the former.

Learning phase To tune the model, prior games from the same participant are assumed to be available. This prior information serve to estimate the influenceability parameters $\alpha_i(1)$ and $\alpha_i(2)$. No prior information is available for Scenario 0 and there is therefore no learning phase in this first case.
In Scenario 1, the influenceability parameters of the participant are estimated independently of the data from other participants (*individual influenceability* method). This is the only feasible method when no prior data on other participants is available. In this case, parameters $\alpha_i(1)$ and $\alpha_i(2)$ are determined by minimizing the mean square error (MSE) (as in equation(4.1)).
In situations where the number of prior games available to estimate the influenceability parameters is small, little information is available on the participant's past behavior. This may result in unreliable parameter estimates. To cope with such situations, another scenario – Scenario 2 – is considered : besides having access to prior data from the targeted participant, part of the remaining participants (half of them, in our study) are used to derive the typical influenceabilities in the population. We decided to use an expectation maximization (EM) algorithm to classify the population into classes of similar influenceabilities [Bishop et al. (2006)]. This algorithm starts with a random partition of individuals into a predefined number of classes[4]. Then, the algorithm successively repeats two steps : 1) for each class C_k, determine the parameters $\alpha^{(k)}(1)$ and $\alpha^{(k)}(2)$ optimizing the likelihood of all the data in the class; 2) affect each individual i to the class C_k for which the parameters result in the highest likelihood. When it comes to choosing the optimal number of classes, we want our model to provide a meaningful characterization of the data (small *bias*) while avoiding overfitting on the training data (small *variance* of the estimator) (Bishop et al., 2006, p168). These typical influenceabilities serve as a pool of candidates. The prior games of the targeted participant are used to determine which candidate yields the smallest mean square error (*population influenceability* method). In other words, we derive from prior games a repartition vector $r_i(t) = k$ that associates to each participant the class that minimize the MSE. The procedure to determine which typical candidate best

[4]In the present study, we analyzed the effect of 1,2,3 and 4 classes on the prediction error

suits a participant requires less knowledge than the one to accurately estimate their influenceability without a prior knowledge on its value.

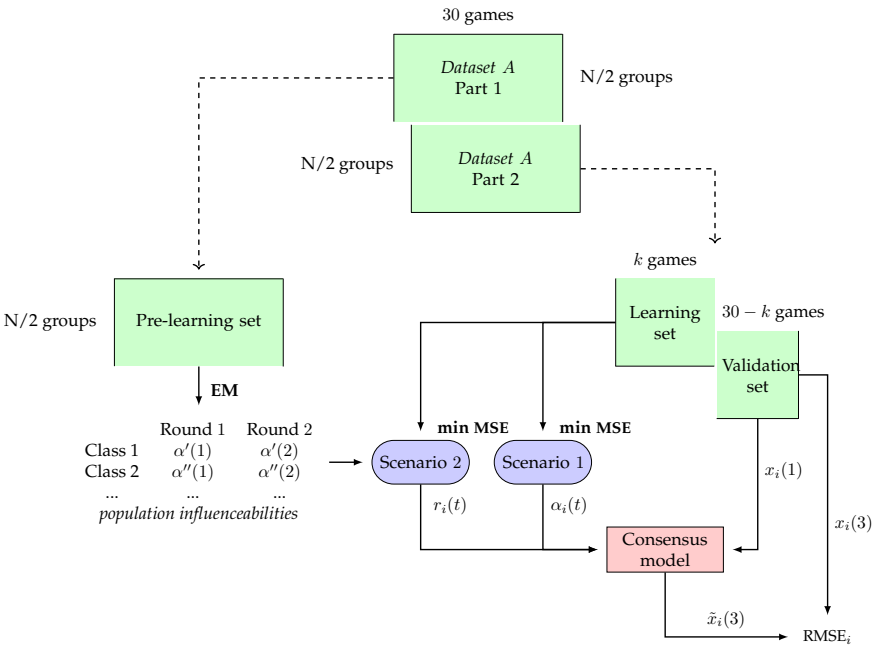


Fig. 4.4 Overview of the crossvalidation process. Splitting is repeatedly performed, at random, until convergence of the global RMSE.

Validation phase The three scenarios are validated via crossvalidation. The validation procedure of the last scenario (i.e., access to prior data from participants, existence and access to typical influenceabilities) starts by randomly splitting the set of participants (N groups of 6 players) into two equal parts, as depicted in Figure 4.4. The prediction focuses on one of the two halves while the remaining half is used to derive the typical influenceabilities in the *population influenceability* method.

In the half serving for prediction, our model is assessed via repeated random sub-sampling crossvalidation : for each participant, a training subset of the games (of varying size k in Figure 4.4) is used to assign the appropriate typical influenceabilities to each participant. The rest of the games serves as the validation set to compute the root mean square error (RMSE) between the observed data and predictions. The error specific to participants i is denoted as RMSE_i. The results are compared for various training set sizes. The RMSE is obtained from averaging errors over 300 iterations using a different randomly

selected training set each time.
To compare Scenario 0 and Scenario 1 with Scenario 2, we only consider the half serving for prediction. Scenario 0 (no data available) does not require any training step. For Scenario 1 (access to prior data from participants), instead of learning affectation of typical influenceabilities to each participant, the learning process directly estimates the influenceabilities $\alpha_i(1)$ and $\alpha_i(2)$ for each participant without prior assumptions on their values. The whole validation process is also carried out reversing the role of the two halves of the population and a global RMSE is computed out of the entire process.

4.4.3 Prediction accuracy

We now compare the *individual influenceability* methods (Scenario 1) with the *population influenceability* methods (Scenario 2) on *Dataset A*. Figure 4.5 first presents the RMSE obtained on validation sets for the *gauging game*. These are computed on the final rounds using the model (4.6-CMOD) with parameters fitted on training data of varying size. All RMSEs were obtained on validation games. The error bars provide 95% confidence intervals for the RMSEs.

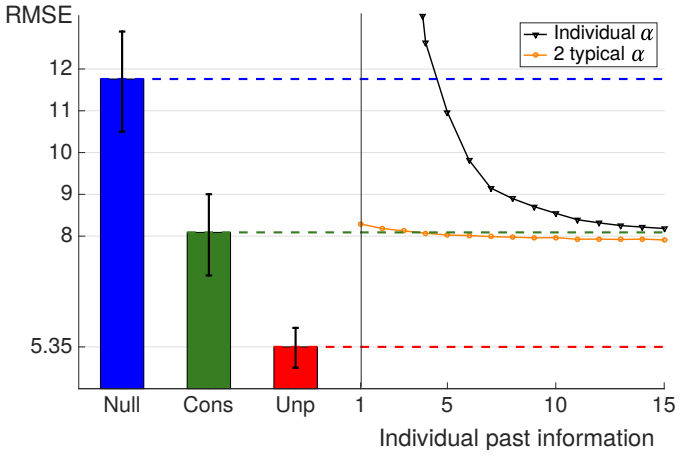


Fig. 4.5 *Gauging game* **- Root mean square error (RMSE) of the predictions for the final round.** The RMSEs are obtained from crossvalidation. The middle horizontal green bar (*Cons*) corresponds to fitting using the same typical couple (= 1 class) of influenceability for the whole population.

The left bar chart displays crossvalidation errors for models that does not depend on the training size. In particular, it includes the null model (top blue horizontal line). As a reminder, the null model assumes constant opinion with

no influence, that is $\alpha(t) = 0$, and consequently does not depend on training set size.

By contrast, the *individual influenceability* method which, for each individual, fits parameters $\alpha_i(1)$ and $\alpha_i(2)$ based on training data, is sensitive to training set size. Its prediction efficiency is displayed by the decreasing black curve (triangle dots) in Figure 4.5. The *individual influenceability* method performs better than the null model when the number of games used for training is higher than 5 but its predictions become poorer otherwise, due to overfitting.

Overfitting is alleviated using the *population influenceability* methods which restrict the choice of $\alpha(1)$ and $\alpha(2)$, making it robust to training size variations. The population method which uses only one typical couple of influenceability as predictor presents one important advantage. It provides a method which does not require any prior knowledge about the participant targeted for prediction. It is thus insensitive to training set size (« *Cons* » in Figure 4.5). This method improves by 31% and 18% the prediction error for the two types of games compared to the null model of constant opinion.

The population methods based on two or more typical couples of influenceability require to possess at least one previous game by the participant to calibrate the model (« *2 typical* α » in Figure 4.5). These methods are more powerful than the former if enough data is available regarding the participant's past behavior (2 or 3 previous games depending on the type of games). The slightly decreasing orange curve (round dots) corresponds to fitting choosing among 2 typical couples of influenceability.

Similar results are observed for the *counting games* (Figure 4.6). The RMSEs have been normalized to the range 0-100 to be comparable with the errors displayed at Figure 4.5. As for the *gauging games*, the *individual influenceability* method becomes more efficient than the null model for training sets higher than 5.

Unpredictability We reported the intrinsic noise estimation (see Chapter 3) by a red dashed line in Figures 4.5 and 4.6. As a reminder, this bottom line corresponds to the amount of prediction error which is due to the intrinsic unpredictability of judgment revision. Theorically, no model can make better predictions than this threshold. Taking the intrinsic unpredictable variation thresholds as a reference, the relative prediction RMSE is more than halved when using the time varying influenceability model (4.6-CMOD) with one couple of typical influenceabilities instead of the null model with constant

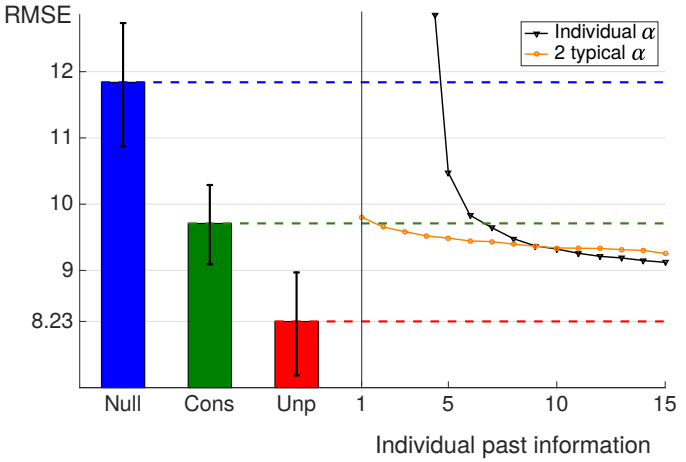


Fig. 4.6 *Counting game* **- Root mean square error (RMSE) of the predictions for the final round.** The RMSEs are obtained from crossvalidation. The middle horizontal green bar (*Cons*) corresponds to fitting using the same typical couple (= 1 class) of influenceability for the whole population.

opinion. This corresponds to an improvement[5] r_3^{imp} of 0.57 for the *gauging game* and 0.59 for the *counting game*.
Also, more than two thirds of the prediction error made by the consensus model is due to the intrinsic unpredictability of the decision revision process, leading to a ratio r_3^{noise} of 0.66 (resp. 0.85) for the *gauging game* (resp. the *counting game*).

Number of classes The number of typical couples of influenceabilities to use depends on the data availability regarding the targeted participant. This is illustrated in Figure 4.7. The modification obtained using more typical influenceabilities for calibration is mild. Moreover, too many typical influenceabilities may lead to poorer predictions due to overfitting. This threshold is reached for 4 couples of influenceabilities in the *gauging game* data. As a consequence, it is advisable to restrict the choice to 2 or 3 couples of influenceabilities. This analysis shows that possessing data from previous participants in a similar task is often critical to obtain robust predictions on judgment revision of a new participant.

Additional remarks The error bars in Figure 4.5 and Figure 4.6 provide 95% confidence intervals for the RMSEs. They confirm statistical significance of the

[5]See last section of Chapter 3.

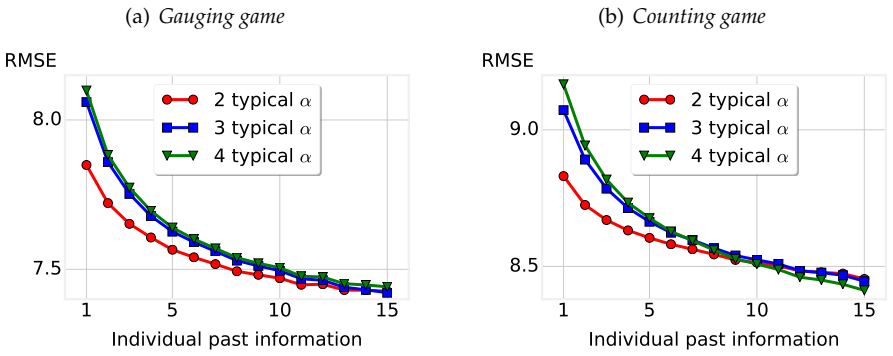


Fig. 4.7 Predictive power of the consensus model when the number of typical couples of influenceability used for model calibration varies. Root mean square error (RMSE) plotted for training set size from 1 to 15 games. For the *counting game*, the RMSE has been scaled by a factor of 5 to be comparable with the *gauging game*. For visualisation purposes, the y-values are displayed on different scale ranges between the two graphs.

difference between RMSEs. For clarity, the error bars are provided only for regression methods which do not depend on training set size. For completeness, Appendix B.1 provides error bars for all models.

RMSEs are used in this study since it corresponds to the quantity being minimized when computing the influenceability parameter α. Alternatively, reporting the *Mean Absolute Errors* (MAEs) may help the reader to obtain more intuition on the level of prediction error. For this reason, MAEs are provided in Appendix B.2.

We remind that the intrinsic variation at round 3 is an underestimation[6] of the intrinsic variation. From a practical point of view, this means that the actual intrinsic unpredictability threshold is even closer to the predictions made by the consensus model than displayed in Figures 4.5 and 4.6. In other words, there is even less space to improve the predictions provided by the consensus model, since more than two thirds of the error comes from the intrinsic variation rather than from the model imperfections.

The methodology was also assessed for second round predictions instead of final round predictions. The results also hold in this alternative case, as described in *Second round predictions* section in Appendix B.3.

[6] See Section 3.4.3 of the previous chapter

4.4.4 Prediction at the individual level

Is the prediction error homogeneous for all influenceability values ? To see this, each bin of the influenceability distributions in Figure 4.8 is colored to reflect average prediction error. The error is given in terms of root mean square error (RMSE). The color corresponds to the prediction error regarding participants having their influenceability falling within the bin. This information shows that the model makes the best predictions for participants with a small but non-negative influenceability. On the contrary, predictions for participants with a high influenceability are less accurate.

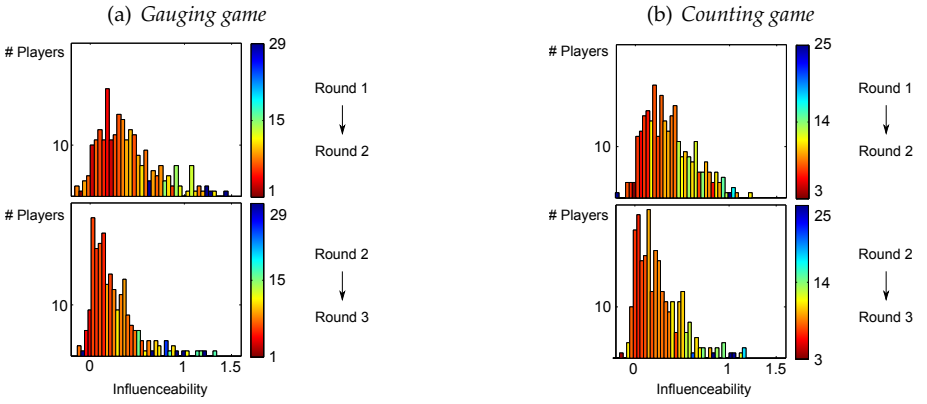


Fig. 4.8 Influenceability with associated predictions. The colormap corresponds to the average prediction RMSE of participants in each bin.

4.5 Practical implications

4.5.1 Empirical consensus

Because of social influence, groups tend to reduce their disagreement. However, this does not necessarily imply that groups reach consensus. To test how much disagreement remains after the social process, the distance between individual judgments and mean judgments in corresponding groups is computed at each round. The results are presented for the *gauging game*. The same conclusions also hold for the *counting game*. Figure 4.9 presents the statistics summary of these distances. The median distances are respectively 5.5, 4.2 and 3.5 for the three successive rounds, leading to a median distance reduction of 24% from round 1 to 2 and 16% from round 2 to 3. In other words, the contraction of opinion diversity is less important between rounds 2 and 3 than between rounds 1 and 2 and more than 50% of the initial opinion diversity is preserved at round 3.

If one goes a step further and assumes that the contraction continues to lessen at the same rate over rounds, it may be that groups will never reach consensus. This phenomenon is quite remarkable since it would explain the absence of consensus without requiring non-linearity in social influence. An experiment involving an important number of rounds would shed light on this question and is left for future work.

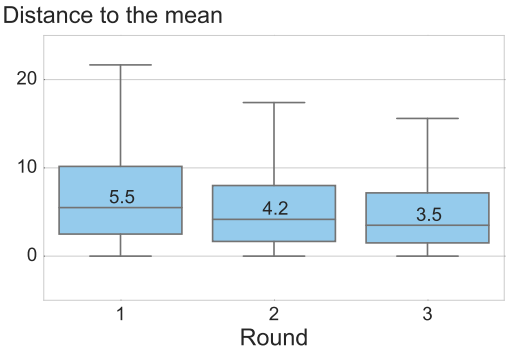


Fig. 4.9 *Gauging game* **: distance of participants' judgment to mean judgment.** Is displayed only the data coming from games where for all 3 rounds at least 5 out of 6 participants provided a judgment. This ensures that the judgment mean and standard deviation can be compared over rounds. The box spreads around the median value q_2 (horizontal line) from the lower quartile (q_1) to the upper quartile (q_3) of the sample. The upper whisker extends below the value $q_3 + 1.5(q_3 - q_1)$ and the lower whisker extends above the value $q_1 - 1.5(q_3 - q_1)$.

4.5.2 Individual and group performance

Each game is characterized by a true value, corresponding to an exact proportion to guess (for *gauging games*) or an exact number of items displayed to the participants (for *counting games*). Whether social influence promotes or undermines individual performance can be measured for the two tasks. Individual performance can also be compared to the group performance.

A measure of performance At each round t, a participant's success is characterized by the root mean square distance to truth, denoted $E_i(t)$. The error $E_i(t)$ depicts how far a participant is to truth. Errors are normalized to fit in the range $0 - 100$ in both types of tasks so as to be comparable. A global measure

of individual errors is defined as the median over participants of $E_i(t)$, and the success variation between two rounds is given by the median value of the differences $E_i(t) - E_i(t+1)$. A positive or negative success variation corresponds respectively to a success improvement or decline of the participants after social interaction.

Wisdom of the crowd As a reminder of Chapter 2, the wisdom of the crowd is a statistical effect stating that averaging over several *independent* judgments yields a more accurate evaluation than most of the individual judgments would. Interestingly, in the first round[7], the wisdom of the crowd effect is more prominent in the *gauging game* than in the *counting game*: the median individual error is 37% higher than the median error of the mean opinion in the *gauging game* while it is only 8% higher[8] in the *counting game*. The reason for this difference can be interpreted from Figure 4.10. The graphs display how opinions are distributed around the true value for the *gauging game* (A) and the *counting game* (B). We observe that most opinions underestimate the true value. However, the bias is more important in the *counting game* which explains – according to the considerations discussed in Chapter 2 – that the wisdom of the crowd is more prominent in the *gauging game* in the first round. This explains the differences between mean opinion errors and individual errors observed in Figure 4.11.

Does social influence promote individual performance ? The errors are displayed in Figure 4.11. The results are first reported for the *gauging game*. The median error $E_i(t)$ for rounds 1, 2 and 3 are respectively 11.9, 10.0 and 9.8 (Figure 4.11-(A)). It reveals an improvement with a success variation of 1.9 and 0.2 for $E_i(1) - E_i(2)$ and $E_i(2) - E_i(3)$ respectively (p-values $< 10^{-5}$, sign test), showing that most of the improvement is made between first and second round. Regarding the *counting game*, the median error $E_i(t)$ for rounds 1, 2 and 3 are respectively 23.1, 22.1 and 21.6 (Figure 4.11-(B)). Note that the errors for the *counting game* have been rescaled by a factor of 5 to fit in the range $0 - 100$. This corresponds to an improvement with a success variation of 1.0 and 0.5 for $E_i(1) - E_i(2)$ and $E_i(2) - E_i(3)$ respectively (the significance of the improvement is confirmed by a sign test with p-values $< 10^{-5}$). Running Example 3 provides an illustrative explanation for the observed improvement.

[7]Only the first round is discussed here because, in the subsequent rounds, the opinions are no longer independent, a criterion required for the wisdom of the crowd to occur.

[8]The hypothesis stating that the mean opinion is closer to the truth than the individual opinion is statistically tested using a Mann–Whitney–Wilcoxon test (p-value < 0.01 for both games).

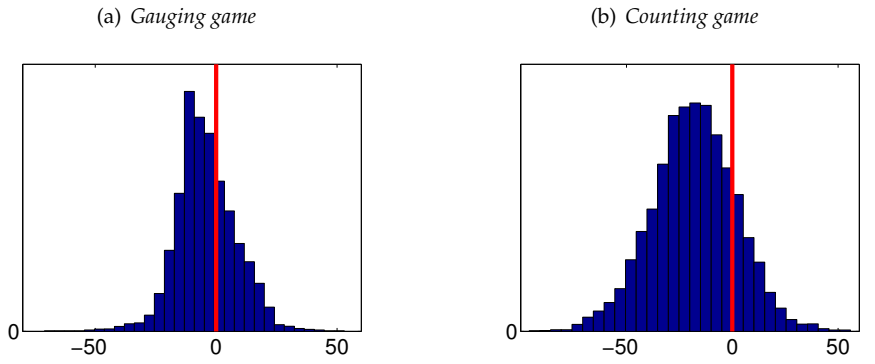


Fig. 4.10 Deviation of opinions to truth during the lone round 1. Red vertical lines split the histogram into the opinions contributing negatively to the distance between mean opinion and truth (left) and the opinions contributing positively to the distance between mean opinion and truth (right).

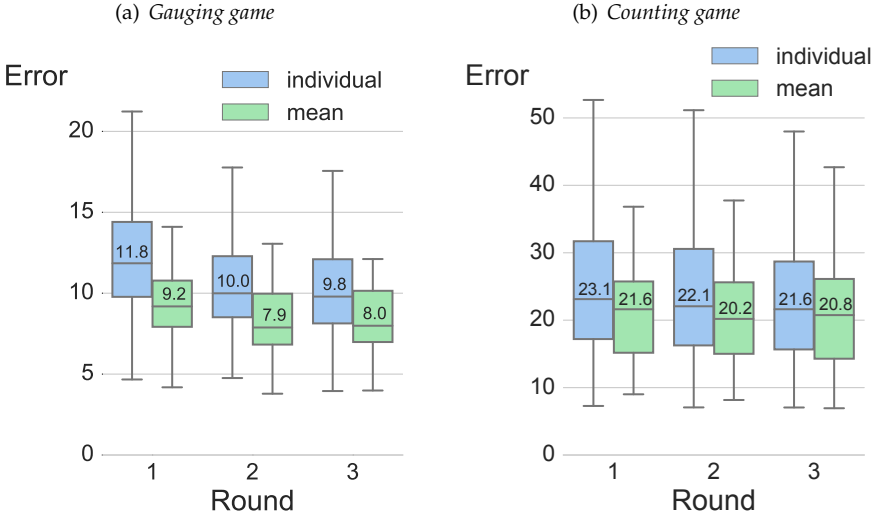
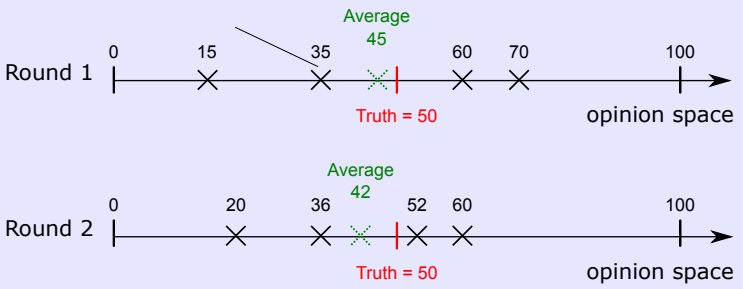


Fig. 4.11 Individual and average performances. Root mean square distance from individual opinions to truth (left) and root mean square distance from mean opinion to truth (right). Values shown for the *counting game* have been scaled by a factor of 5.

Running example 3 - Understanding the individual improvement

The reason why social influence helps the individual performance is a combination of two factors. First, at round 1, the wisdom of the crowd effect specifies that the mean opinion is closer to the truth than the individual opinion is. Second, in accordance to the consensus model (4.6-CMOD), individuals move closer to the mean opinion over subsequent rounds.

To illustrate, we design an example of a *gauging game* with interaction over two rounds amongst four players :



At round 1, we observe that the mean opinion (45) is closer to truth (50) than the individual opinions, which is a direct consequence of the wisdom of the crowd effect. The fact that the mean opinion is more accurate than individual opinions leads to posit that participants using the mean opinion to form their own opinion, *i.e.*, those with higher influenceability α_i, will increase their performance. This is also reflected in the running example : when the players are attracted towards the mean, their average error drops from 20 to 14. However, we note that the interaction does not especially promote the aggregate performance. In this case, the aggregate opinion becomes even worse, moving from 45 to 42.

Back to *Dataset A*, we examine relationships between success variation and the model parameters α_i by computing partial Pearson correlations ρ controlling for the effect of the rest of the variables[9]. Influenceability $\alpha_i(1)$ between round 1 and 2 and improvement are positively related with $\rho(\alpha_i(1), E_i(1) - E_i(2)) = 0.41/0.22$. This is in accordance to the posited hypothesis.

[9]Only significant Pearson correlations are mentioned (p-value < 0.05). Pearson correlations are given by pairs : the first value corresponds to the *gauging game* while the second value corresponds to the *counting game*.

The *wisdom of the crowd effect* found at round 1 implies that participants who improve more from round 1 to 2 are those who give more weight to the average judgment. Since the wisdom of the crowd effect is more prominent in the *gauging game* than in the *counting game*, it is consistent that the correlation is higher in the former than in the latter. A similar effect relates success improvement and the influenceability $\alpha_i(2)$ between round 2 and 3 with $\rho(\alpha_i(2)\,,\, E_i(2) - E_i(3)) = 0.36/0.32$. As may be expected, higher initial success leaves less room for improvement in subsequent rounds, which explains that $\rho(E_i(1), E_i(1) - E_i(2)) = 0.68/0.21$ and $\rho(E_i(1), E_i(2) - E_i(3)) = 0.25/0.16$ (where 0.16 is significant for p-value < 0.01). This also means that initially better participants are not better than average at using external judgments.

Does social influence promote the aggregate performance ? Figure 4.11 also reports the aggregate performance in terms of the root mean square distance from mean opinions to truth in each group of 6 participants. Unlike individual performance, the median aggregate performance does not consistently improve. Regarding the *gauging game*, the median aggregate error is significantly higher in round 1 than in rounds 2 and 3 (p-value $< 10^{-4}$, sign test) and this difference is not significant between rounds 2 and 3 (p-value > 0.05). Regarding the *counting game*, no significant difference is found among the 3 rounds for the median aggregate error (p-value > 0.05). Since the aggregate performance does not consistently improve over rounds, it can be said that social influence does not consistently promote the wisdom of the crowd.

As a consequence, social influence consistently helps the individual performance but does not consistently promote the aggregate performance.

4.5.3 Task influence

The assessment of the predictive power of model (4.6-CMOD) on both types of games provides a generalizability test of the prediction method. The two types of games vary in difficulty. The root mean square relative distance $E_i(1)$ between a participant's first round judgment and truth is taken as the measure of inaccuracy for each participant. The median inaccuracy for the *counting game* is 23.1 while it is 11.8 for the *gauging game*. Mood's median test supports the rejection of equal median with p-value < 0.05. The accuracy of model (4.6-CMOD) is compared for the two datasets in Figure 4.5 and Figure 4.6. Interestingly, the model prediction ranks remains largely unchanged for the two types of games. Similarly, if we come back to Figure 4.1, influenceability distributions do not vary significantly between the two games. A two-sample Kolmogorov-Smirnov test fails to reject the equality of distribution null hypothesis of equal

median with p-value > 0.65 and a KS distance of 0.06 for both $\alpha_i(1)$ and $\alpha_i(2)$. This means that although the participants have an increased difficulty when facing the *counting game*, they do not significantly modify how much they take judgments from others into account. Additionally, the relationships between participants' success and influenceability are preserved for both types of games. The preserved tendencies corroborate the overall resemblance of behaviors across the two types of games. These similarities indicate that the model can be applied to various types of games with different levels of difficulty.

4.6 Extending the consensus model

In this final section, we investigate two ways of extending the consensus model (4.6-CMOD) for further crowdsourcing experiments. The first extension depicts a case where more than three rounds compose the interaction process. The second variation proposes to analyze the effect of introducing nonlinearities.

4.6.1 Additional rounds

It is reasonable to assume that the opinion of individuals settles after some time even though disagreement remains [Mason, Conrey, and Smith (2007)]. When we are faced with a revision process including more than 3-rounds ($R > 3$), we can model this using a time-decaying confidence of the form $\gamma_i e^{-\beta_i t}$. We assume that the confidence decays exponentially in time. This leads to two parameters per person, that is, the initial confidence and its decay rate:

$$\hat{x}_i(t+1) = x_i(t) + \gamma_i e^{-\beta_i t}(\bar{x}(t) - x_i(t)).$$

The solution to the log likelihood maximization may not be unique and belongs to the solutions of the following coupled equations:

$$\begin{cases} \gamma_i = \dfrac{\sum_{\text{pictures}} \sum_{t=1}^{R-1} e^{-\beta_i t}(X_i^p(t+1) - X_i^p(t)) \cdot (\bar{X}^p(t) - X_i^p(t))}{\sum_{\text{pictures}} \sum_{t=1}^{T-1} e^{-2\beta_i t}(\bar{X}^p(t) - X_i^p(t))^2} \\ \gamma_i = \dfrac{\sum_{\text{pictures}} \sum_{t=1}^{R-1} te^{-\beta_i t}\,(X_i^p(t+1) - X_i^p(t)) \cdot (\bar{X}^p(t) - X_i^p(t))}{\sum_{\text{pictures}} \sum_{t=1}^{T-1} te^{-2\beta_i t}(\bar{X}^p(t) - X_i^p(t))^2} \end{cases}.$$

In this case, the best parameters are found by computing the highest values of the log likelihood for each couple of parameters (γ_i, β_i) satisfying the previous equations.

4.6.2 Introducing nonlinearities

The consensus model (4.6-CMOD) assumes that the opinion change $x_i(t+1) - x_i(t)$ grows linearly with the distance between $x_i(t)$ and the mean opinion $\bar{x}(t)$. Figure 4.12 displays the evolution $x_i(t+1) - x_i(t)$ against the distance to the mean $\bar{x}(t) - x_i(t)$. The color in each cell corresponds to the number of data points falling in the cell, and the straight black line represents the linear regression $x_i(t+1) - x_i(t) = \beta_0 + \alpha(t)(\bar{x}(t) - x_i(t))$.

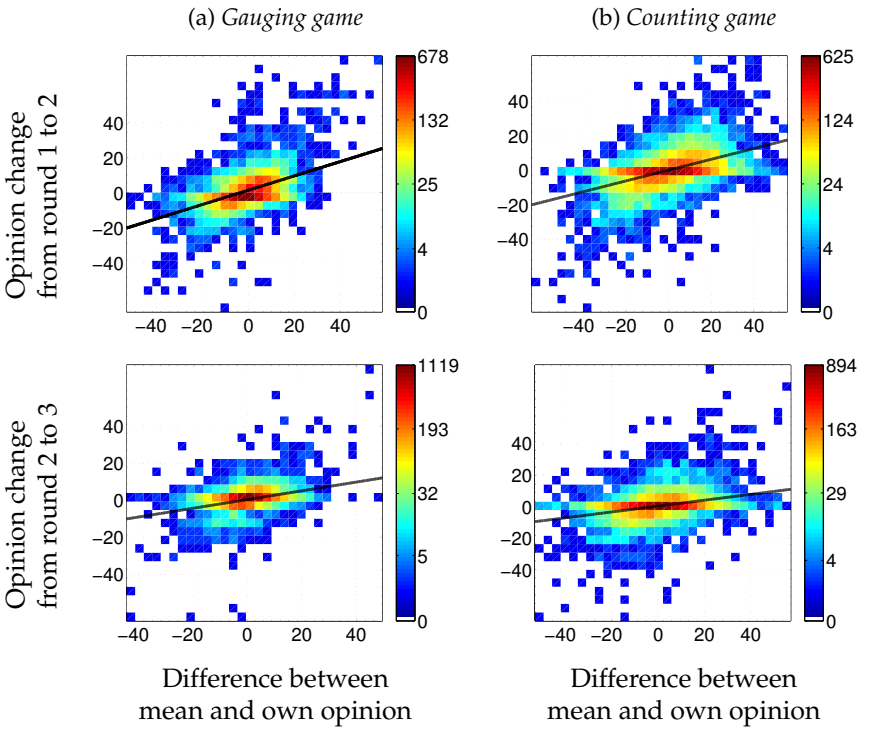


Fig. 4.12 Opinion change $[x_i(t+1) - x_i(t)]$ versus difference between mean and individual opinion $[\bar{x}(t) - x_i(t)]$. The color scale is logarithmic.

The linear assumption is tested[10] against the alternative

[10]The following numerical statistics values are reported for the opinion change between rounds 1 and 2 for the *gauging game*; the same conclusions hold for the *counting game* and for the opinion change between rounds 2 and 3

$$x_i(t+1) - x_i(t) = \beta_0 + \alpha_1(\bar{x}(t) - x_i(t))^\gamma$$

with $\gamma \neq 1$. The linearity test provided in [Niermann (2007)] applied to our data gives a statistics $P = -1.4 \times 10^7$ with empirical variance $\text{var}(P) = 4 \times 10^{14}$ so that we fail to reject the null hypothesis $\gamma = 1$ (p-value $= 0.5$).

4.7 Concluding remarks

The way online social systems are designed has an important effect on judgment outcome [Muchnik, Aral, and Taylor (2013)]. Operating or acting on these online social systems provides a way to significantly impact our markets, politics [Bond, Fariss, Jones et al. (2012)] and health. Understanding the social mechanisms underlying opinion revision is critical to plan successful interventions in social networks. It will help to promote the adoption of innovative behaviors (e.g., quit smoking [Christakis and Fowler (2008)] or eat healthy [Valente (2012)]).

The present chapter shows that it is possible to model opinion evolution in the context of social influence in a predictive way. When the data regarding a new participant is available, parameters best representing their influenceability are derived using mean-square minimization. When the data is scarce, the data from previous participants is used to predict how the new participant will revise their judgments. To validate our method, results were compared for two types of games varying in difficulty. The model performs similarly in the two experiments, indicating that our influenceability model can be applied to other situations.

The decaying influenceability model after being fit to the data suggests that despite opinion settlement, consensus will not be reached within groups and disagreement will remain. This suggests that there needs to be incentives for a group to reach a consensus. The analysis also reveals that participants who improve more are those with highest influenceability, this independently of their initial success.

The design and validation of models of opinion revision will enable to create a bridge between system engineering and network science. The consensus model can serve as a building block to more complex models when collective judgments rely on additional information exchange.

Chapter 5

Uncovering Political Ideologies from Digital Traces

When participants take part in social influence experiments, their opinions are usually measured by inviting the participants to directly express their views on specific topics. Repeatedly conducting large scale studies and time-dependent surveys becomes quickly intractable and very costly. Since individual opinions involved in real world processes are not directly available, researchers had to calibrate their models on global measures such as levels of controversy [Iñiguez, Török, Yasseri et al. (2014)] or distribution of votes [Caruso and Castorina (2005); Fortunato and Castellano (2007)].

The study carried out in this chapter investigates the particular case of public opinion in political science. A contentious issue in the public opinion literature relates to citizens ' ability to deal with political abstract concepts. In order to understand trends and shifts within a population of interest, public opinion research often requires recurrent, comparable, and valid estimates of political positions, defined by the notion of *ideology*. Reference to left-right ideology is prominent in everyday political discussions. These terms provide handy yardsticks to compare or summarize politicians' or citizens' views on policies.

In the last years, scholars started to look at more scalable data sources in order to widen the scope of public opinion research. The rise of social media offers new sources of information that are publicly available for different actors on a large scale. Social media can allow researchers to perform cross-national and real time public opinion analysis. However, many doubts persist in the academic community regarding the validity of such data for studying public

opinion. Can these new sources of information be used to capture ideologies? If so, are these ideologies comparable to those measured in surveys?

In the following, we present methodologies for measuring ideology on social media. Estimates are validated using a unique large-scale survey of Twitter users collected by Vox Pop Labs[1]. This survey collects the opinions of respondents over political issues from which left-right positions are derived and allow direct comparison with the ideological positions estimated with new social media methods.

We start our work by describing conventional methods to estimate ideologies from various types of information (speeches, roll-call votes,...). We then focus on the extension of these methods to social media data. The core of this chapter is based on the estimation of ideologies from online connections – *network* data – and published comments – *textual* data. We validate the proposed method by showing it highly correlates with existing measures of ideology for different political contexts and for different languages. To demonstrate the usefulness of this method, we show the ability of the derived ideological positions to predict out-of-sample individual-level voting intentions. Interestingly, the predictive power of ideological positions derived from social media can outperform surveys when combining textual with network data.

5.1 From opinions to ideology

Are citizens able to clearly understand the abstract notions behind political concepts? The debate goes back to the origins of survey research when it was found that most citizens lack ideological constraints [Campbell, Converse, Miller et al. (1960); Converse (1964)]. That is, knowing the opinion of an individual over a single political issue gives no information on his preferences over any other political issue. Further research consolidated these findings by revealing low levels of political sophistication and information among citizens [Carpini and Keeter (1997); Luskin (1990)]. However, other scholars show that a clear public "signal" can be obtained by aggregating individual preferences and multi-item scales [Ansolabehere, Rodden, and Snyder (2008); Page and Shapiro (2010); Stimson (2004)]. These works emphasize the importance of being able to position and compare agents on a common scale. They show that with proper measurement it is possible to derive stable positions in the population, and that these are not restricted to political elites [Jessee (2016)]. These positions or ideal points are referred to as *ideologies*.

[1]For more details see: http://voxpoplabs.com

There exists no consensus on a precise definition of ideology [Gerring (1997)]. Yet, few would disagree with a minimalist definition as a belief system—a coherent and consistent view towards the political world—that structures citizens' political preferences [Converse (1964); Lane (1962)]. In addition to the definitional challenges, measurement of ideologies are also complicated by the fact that this abstract concept is not directly observable. The measurement of ideological positions has attracted academic scrutiny for many decades and in various disciplines such as political science, psychology, and sociology [Bonica (2013); Colic-Peisker (2016); Feldman and Johnston (2014)]. Different methods exist and are still being developed to measure these underlying structures [Barberá (2015); Bonica (2013); Poole and Rosenthal (1985)]. In most modern democracies, a unidimensional left-right scale is often used to summarize or describe parties' and citizens' political views [Bonica (2014)]. These ideologies are not simply descriptive handy terms. In fact, ideologies are seen by many as a major explanatory factor underlying voting behavior [Conover and Feldman (1981)]. Spatial theory's basic idea, for example, is that citizens' and parties' behavior is determined by their relative positions in a given ideological landscape [Davis, Hinich, and Ordeshook (1970); Downs (1957); Jessee (2012)].[2]

From a mathematical perspective, ideologies are usually associated to a specific number of dimensions. In the following, these are noted ϕ and can be represented in the political space by a d-dimensional vector for every agent i, for example, $\phi_i \in \mathbb{R}^d$. Extracting these dimensions is called ideological scaling. Intuitively, we consider that two members who share similar political patterns or opinions are supposed to be close in the ideological space.[3]

In this chapter, we attempt to estimate the ideological positions of social media users from their online activity. For this purpose, we assume for simplicity constant and unidimensional ($d = 1$) ideologies. Time independence is justifiable on the grounds that the analysis is performed on short periods of times. In addition, most scaling methods described below can easily be extended to a multidimensional analysis.

[2] Of course, the state of the literature on spatial theory is far more complex. One can simply mention the prominence of the proximity-directional debate in that literature [Rabinowitz and Macdonald (1989)]. Or the role of valence issues in contrast to positional issues in these models [Stokes (1963)].

[3] In the present, we consider the ideological space as a Euclidean space.

5.2 Conventional measurement methods

Public opinion research often measures ideological positions (ϕ) based on survey data. They generally consist in aggregating respondents' opinions over multiple survey questions regarding topics that might contribute to the latent ideological dimension. This aggregation is performed using additive scales or dimensionality reduction methods [Ansolabehere, Rodden, and Snyder (2008); Ellis and Stimson (2012)]. Other methods focused on political candidates and scale ideological positions based on roll-call votes, political speeches, or party manifestos [Monroe and Maeda (2004); Poole and Rosenthal (1985); Slapin and Proksch (2008)].

Since estimation methods are usually scale invariant, the measured positions $\hat{\phi}$ are often rescaled or standardized for identification purposes. In this work, we decided to set the mean of the estimates to 0 and their standard deviation to 1. Even after rescaling, the ideological axes can still be reversed (which is called *reflexion invariance*). We handle reflexion indeterminacies by associating positive values to conservative parties and, conversely, negative ideologies to liberal parties.

5.2.1 Policy Preferences

One of the conventional extraction methods consists in measuring ideologies from expressed policy preferences towards political issues. Ansolabehere, Rodden, and Snyder (2008) show that multi-item scales based on preferences over political issues should be preferred to single ideological self-placement questions as these considerably reduce the measurement error.

Identifying patterns of opinions from multiple measurement is usually performed by factor analysis [Rummel (1967)]. We denote $x_{ik} \in \{1, 2, \dots, L\}$ the expressed opinion of a respondent i related to a specific issue k in a survey (e.g., on a Likert scale[4]). The related factor analysis model (5.1-PREF) assumes one factor (i.e., the citizen's ideology ϕ_i) and independent error terms ϵ_k:

$$x_{ik} = \theta_k \phi_i + \epsilon_k \,, \quad i = 1 \dots n \tag{5.1-PREF}$$

where the terms θ_k weigh the importance of each question in reflecting the underlying ideology. These are called factor loadings and map the ideological space to the observed opinions. Parameters are generally computed by maxi-

[4]The Vox Pop Labs Survey use $L = 5$ in the present study.

mum likelihood (ML) methods.

Nonetheless, repeatedly obtaining survey data about political preferences for different actors, on a large scale, across countries, and over time is often cost prohibitive [Bond and Messing (2015); Slapin and Proksch (2008)].

5.2.2 Roll-Call Votes

Political candidates' ideological positions have been studied under different angles during the last few decades and encouraged the development of different methodologies. A popular method of scaling politicians aims at estimating ideologies based on voting decisions in deliberative assemblies. In those sessions, referred to as *roll-call* sessions, each candidate is given a list of decisions and decides whether to vote for or against a proposal. Votes are recorded and can be represented by a binary matrix Y whose elements consist of the affirmative ($y_{jk} = \text{Yes}$) and negative ($y_{jk} = \text{No}$) votes.

Ideologies are then estimated by applying multidimensional scaling techniques [Kruskal and Wish (1978)]. These methods model each legislator's utility for a policy and the resulting estimated position depends on the choice of a specific utility function. Two common functions are the Gaussian utilities introduced in the DW-NOMINATE applications [Poole and Rosenthal (2000)] and the quadratic functions [Clinton, Jackman, and Rivers (2004)]. The DW-NOMINATE method is by far the leading scaling process when analyzing roll-call data. The acronym stands for dynamic, weighted, nominal three-step estimation. In the case of constant and unidimensional ideologies, the model takes the following form :

$$\begin{aligned} Pr(y_{jk} = \text{Yes}) &= \Phi\left(\beta \Delta U\right) \\ \text{with } \Delta U &= \exp[-0.5(\phi_j - \theta_k^{\text{Yes}})^2] \\ &\quad - \exp[-0.5(\phi_j - \theta_k^{\text{No}})^2] \end{aligned} \tag{5.2-ROLL}$$

and expresses the probability that a legislator vote *yes* to a decision k. Notation Φ stands for the standard normal cumulative distribution function and β is a scaling parameter. The variable θ_k^{Yes} (resp. θ_k^{No}) represents the position in the ideological space for a positive (resp. negative) sentiment towards issue k. Intuitively, expression (5.2-ROLL) models the legislator's gain ΔU (or loss) to vote a yea instead of a nay during the roll-call vote session.

When looking at ideological positions, roll-call data is considered an estab-

lished standard to estimate ideological positions of political elites [Carroll, Lewis, Lo et al. (2009); Poole and Rosenthal (1987, 2001)]. Nonetheless, this data source comes with some restrictions. For instance, the data is only available for elected officials and the models can only scale positions in political contexts with low levels of party discipline. These constraints make the American context a top choice for most studies interested in political ideologies. However, the scope of political behavior is not limited to political actors nor the United States. For instance, if we look at Canada, where party discipline is high [Kam (2009)], roll-call models would not be as effective at providing individual politician positions in the ideological landscape [Godbout and Høyland (2011)].

For these reasons, it is important to explore new data sources and develop methods that give us more flexibility and allow us to estimate ideological positions not only for a broader set of political actors, but also for a broader set of political contexts.

5.2.3 Speeches and manifestos

Elected officials and political parties' ideological positions can be estimated from their political speeches and party manifestos. The assumption made in the content extraction process suggests that the use of specific political terms reflects the authors' opinions as these select specific combination of terms. These terms can be expressed by a politician or a political party multiple times.

Monroe and Maeda (2004) proposed a scaling model that extends the roll-call analysis to textual contents. Their work considers a Poisson process to model the use of words in speeches and documents. The process first builds a matrix Z, also called a term-document matrix (TDM), where each element z_{jw} corresponds to the number of times the word w is present in party j manifesto (see Running Example 4 for further details). We then express the expected value of z_{jw} as a function of the ideologies (ϕ_j):

$$\begin{aligned} Pr(z_{jw} = k) &= f(k; \lambda_{jw}) \\ \text{with } \lambda_{jw} &= \exp\left(s_j + p_w + \theta_w \, \phi_j\right) \end{aligned} \tag{5.3-WORD}$$

where $f(k; \lambda_{jw})$ represents the Poisson probability mass function. The θ_w parameters take into account the impact of words in the author's position and have a similar function as the factor loadings of equation (5.1-PREF). The model introduces two additional factors: the words' popularity (p_w) and the political elites' susceptibility (s_j) to deal respectively with more frequent words and longer manifestos or speeches.

Running example 4 - The term-document matrix

Term-document matrices (TDMs) are built from a list of tuples (author - document) where the documents are a collection of sentences. As an example, let us consider the following list:

- D. Trump - *Building a wall, letting them pay for the wall.*
- B. Sanders - *Combating climate change.*
- H. Clinton - *Taking on the threat of climate change.*

The order of words and sentences is neglected since TDMs do not incorporate time dependence. Traditionally, researchers first apply a series of filters to improve the scaling quality. For instance, algorithms can discard stop words and merge words that are closely related (e.g, the words *fishing* and *fisher*) into one single root (e.g., *fish*). This is called *stemming*. These operations performed on the above example lead to the following list of stems:

build, wall, let, pay, combat, climat, chang, take, threat

The algorithm then associates identification numbers $(1, \ldots, 3)$ to the authors (here, politicians). Similarly, identifiers $(1, \ldots, 9)$ are associated to the words. A rectangular matrix Z of size 3×9 is built where each element z_{iw} tallies the occurrences of a word w that appear in an author's document i:

Politicians	
1	D. Trump
2	B. Sanders
3	H. Clinton

Words	
1	wall
2	chang
…	…
9	climat

$$Z = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ \ldots & \ldots & \ldots \\ 0 & 1 & 1 \end{pmatrix}$$

Politicians (columns) — **Words** (rows)

The textual extractions aims at deriving ideal positions from the TDM so that the authors expressing similar ideas have close ideologies in the political space.

Estimators of the unknown parameters are derived using a maximum likelihood (ML) estimation. An iterative approach via expectation maximization (EM) is implemented in the popular *Wordfish* algorithm [Slapin and Proksch (2008)].

5.3 Measuring ideology on social media

Generally, various data are attached to online user accounts, such as the set of their followers or friends, the content they liked, shared or even published. The difficulty to validate estimates based on these social media sources can explain a large part of the skepticism about their scientific value. This task is difficult due to the absence of reliable individual-level information about the users such as their socio-demographic information and formally expressed attitudes over political issues. Therefore, large-scale and recurrent measures of ideologies focus most of the time on political elites [Bonica (2013); Martin, Saalfeld, Strøm et al. (2014); Slapin and Proksch (2008)]. This is problematic as political elites are known to hold more ideologically consistent policy preferences compared to citizens [Jewitt and Goren (2016); Kritzer (1978)].

Yet, research using social media data flourished in the last years perhaps due to the perceived potential of this bulk of information made publicly available by political elites and ordinary citizens. In fact, with proper methods, the information available on social media can be transformed into survey-like data that can enable the study of public opinion at the individual level and also in real time [Beauchamp (2016)].

5.3.1 Mapping surveys and online datasets

We investigate three different elections : the *41st Quebec general election* (from 5 March 2014 to 7 April 2014), the *51st New Zealand general election* (from 14 August 2014 to 20 September 2014) and the *42nd Canadian federal election* (from 4 August 2015 to 20 September 2015).

Our work is based on a large online survey of Twitter users ($n = 62,430$) who decided to share a valid Twitter account for research purposes. This survey, realized by *Vox Pop Labs* (VPL), is actually a subsample of a larger online survey ($n > 3,120,000$) collected through the voting engagement application Vote Compass[5]. In addition to their Twitter account, the participants provided an-

[5]For more details see: http://voxpoplabs.com/votecompass/methodology.pdf

swers to a series of questions, including their vote intention and their opinion on 30 political issues. The question probing participants' vote intention indicates whether they intend to vote and for which party they intend to vote for. The set of 30 political questions are composed of different questions regarding the election context. They are also asked to each party representative. Three examples of questions for the Canadian context are provided in Table 5.1

Canada 2015	
Issue ID	Question
6	How much should be done to accommodate religious minorities in Canada?
12	To what extent should law enforcement be able to monitor the online activity of Canadians?
18	How involved should the Canadian military be in the fight against ISIS?

Table 5.1 Subsample of the question list designed by Vox Pop Labs for Canada 2015

Appendix C.2 provide the complete list of questions as well as details about selection of the issues by the political experts at VPL. As a validation procedure, VPL places all parties on each of the issue using content analysis of party manifestos and other publicly available texts.

Publicly available networks and textual data are collected using Twitter's REST API. Network information is derived from the ability of members to follow one or multiple Twitter accounts. Any account is linked to a list of followers and a list of followees (also denoted as friends). The networks are handled as directed graphs and were collected on the 29th October 2015.[6] On the other hand, textual information is derived from users' statuses – often referred as tweets – and are collected during the campaign periods. These include hashtags and retweets. Only tweets published during the election periods are considered. Networks and tweets are retrieved for citizens and political elites that did not restricted the use of their Twitter account.

5.3.2 The three election contexts

We first describe each context by realizing a policy preferences analysis. Ideological dimensions are obtained using a bidimensional[7] factor analysis on the 30 political issues collected from the VPL surveys.

[6] Twitter REST API does not allow to retrieve past network data.

[7] Equation (5.1-PREF) translates a unidimensional version of the factor analysis. See for instance [Harman (1960)] for the multidimensional extension ($d > 1$).

Canada 2015 Canada has two official languages with a very high level of party discipline. The multipartisan and multilingual characteristics of Canada makes it a challenging case for the application of text-based ideological scaling techniques. The party system in Canada is composed of four national-wide parties, from left to right: the *Green Party of Canada (GPC)*, the *New Democratic party (NDP)*, the *Liberal Party of Canada (LPC)*, and the *Conservative Party of Canada (CPC)*. The last two parties are the only parties to have held office in recent Canadian history. The CPC that won a majority of seats in the Canadian Parliament since 2006 lost to the LPC in 2015.

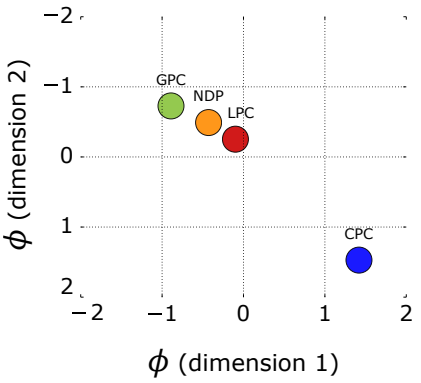


Fig. 5.1 Canada – Parties' positions from policy preferences.

New Zealand 2014 (NZ) The New Zealand party system is very diverse counting seven parties having at least a seat in Parliament. For this analysis, we only retained four of them[8]. The *Green Party of Aotearoa New Zealand* (GRN) and the *Internet Mana* (MNA) both have the particularity of being led by two co-leaders and are clearly characterized by a left-wing political structure. On the opposite, the *New Zealand National Party* (NP) has a center-right position and won nearly 50% of the seats in 2014. The *New Zealand Labour Party* (LAB) is a center-left party and is classically the major opponent of the National Party.

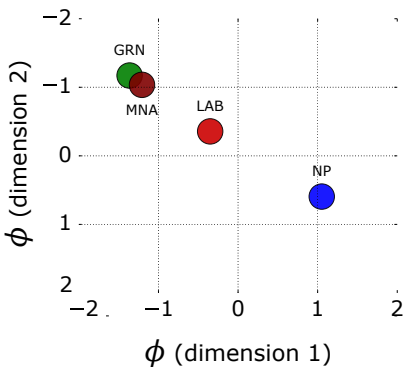


Fig. 5.2 NZ – Parties' positions from policy preferences.

[8]The three remaining parties did not pass the required activity criterion described in section Elites and Citizens.

Québec 2014 The particularity of the Quebec context holds to the impact of the issues related to nationalism and identity on the ideological landscape. There exists on these issues a clear cleavage between political parties and citizens. On the most nationalist side of this spectrum, three parties advocate for Quebec independence: *Option Nationale* (ON), the *Parti Québécois* (PQ), and *Québec Solidaire* (QS). The least nationalist party and most opposed to Quebec independence is the *Quebec Liberal Party (QLP)*, the party that won office in 2014 putting a term to a short-lived PQ minority government. The *Coalition Avenir Québec* (CAQ) positions itself in-between these two poles by advocating more autonomy for the French-speaking province while remaining a member of the Canadian federation. The socio-economic left-right ideological dimension is also structuring Quebec party competition. This dimension follows roughly the same order as the identity dimension with most nationalist parties leaning generally more on the left side of the spectrum relative to federalist parties.

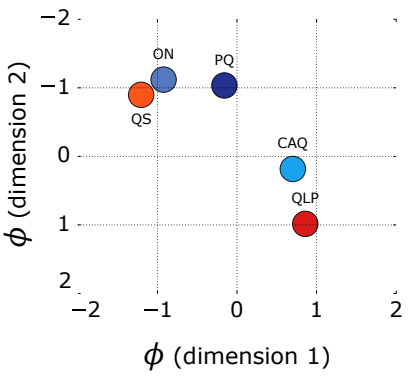


Fig. 5.3 Québec – Parties' positions from policy preferences.

5.3.3 Elites and Citizens

This analysis makes the distinction between two types of social media users: political actors (referred as elites or candidates) and survey respondents (denoted citizens). Political elites are analyzed separately as their political views can be easily associated with those of the political parties they are running for.

Elites The term *elite* refer to political candidates running for office. Their average position in the ideological space is publicly known. Even if some variability in the political elites' ideological positions is expected, it is reasonable to assume that most political elites should be generally closer to their own party than to any other party.

These elites' accounts were identified by performing an automatic search from official candidates' lists through the Twitter REST API and by manually selecting consistent accounts regarding Twitter features (account description and profile pictures). We identify respectively 759, 297 and 131 elites for Canada

2015, New-Zealand 2014 and Quebec 2014. The present work is focused here on Twitter information and therefore on politically active Twitter political elites. Political elites that do not exchange enough political content are automatically discarded. In order to avoid any pre-assumption, a bigram dictionary containing political terms is constructed for each election using an unsupervised process (the creation process is later described in Section 5.3.7). A content filter with threshold $\beta = 5\%$ ignores candidates whose tweets does not include at least 5% of these political terms. Analogously, we require the political elites to appear politically popular on the social network. A second filter eliminates political elites that does not count 25 of the survey citizens as their followees. We finally discarded parties who did not count at least three politicians at the end of this filtering process. Discarding these minor parties from the analysis prevents unreliable conclusions (e.g., small sample sizes for classification). The number of active political elites belonging to a major party after filtering represents a total of $m = 120$ for Canada 2015, $m = 56$ for New-Zealand 2014 and $m = 106$ for Quebec 2014.

Citizens By contrast, citizens refer to survey respondents that are not candidates for the election and represent potential voters[9]. The Vox Pop Labs survey registered $n = 42,634$ (Canada 2015), $n = 11,344$ (Quebec 2014) and $n = 8452$ (New-Zealand 2014) citizens who decided to share a valid Twitter account.
As for the candidates, our analysis takes as its framework only citizens that express political opinions on Twitter. We only keep Twitter users who published content including more than 5% of the political terms[10] during the electoral campaign and who follow at least 3 active candidates. This filtering condition is very constraining since most of the citizens do not express political opinions on the social network, and our analysis is therefore only restricted to politically engaged citizens. The citizens sample sizes after these filtering conditions are $n = 1717$ (Canada 2015), $n = 123$ (New-Zealand 2014) and $n = 796$ (Quebec 2014).

5.3.4 Network Estimates

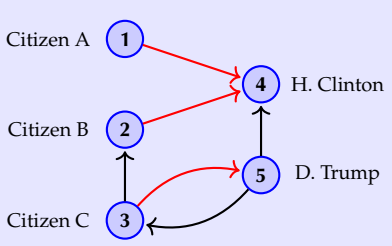
The network methods were first developed from *likes* on Facebook public pages [Bond and Messing (2015)] as well as from *followers* on Twitter politicians accounts [Barberá (2015)]. The ideological estimates based on network information are computed from the existing connections between citizens and political elites on social networks. In our case, these connections are created each time a citizen decides to follow a political elite. The input in the network scaling

[9] Voting is not compulsory for the three considered elections
[10] Political terms refers to those included in the political dictionary.

process is a binary matrix $Y \in \mathbb{R}^{n \times m}$, in which the elements y_{ij} of Y are equal to 1 when a citizen i is connected to a political elite j, otherwise they equal to 0. We call Y the network's *homophilic* matrix (see Running Example 5 for further details).

Running example 5 - The homophilic matrix

The network extraction method only focuses on the connections extending from citizens to political elites. These connections appear when a citizen decides to follow a political figure. As an example, let us consider the following network (a):



(a)

follow → Citizens Elites

	Citizens			Elites	
Citizens	0	0	0	1	0
	0	0	0	1	0
	0	1	0	0	1
Elites	0	0	0	0	0
	0	0	1	1	0

(b)

The network includes 5 agents (3 nodes representing citizen accounts and 2 nodes standing for political elites). Edges connect the followers to the followees (citizen C follows citizen B which in turn follows H. Clinton). The resulting network is a digraph that can be represented by an asymmetric adjacency matrix (b). The matrix citizen labels $(1, 2, 3)$ are intentionally lower than the political elites identifiers $(4, 5)$: the resulting partitioned matrix is made of 4 blocks and its upper right block is called the homophilic matrix. The matrix block associates a list of followers to each one of the political elites. These connections appear as red arrows in the digraph.

The extraction process relies on the assumption that these virtual connections are homophilic in nature. That is, a citizen i tends to follow political elite j when their respective ideologies (ϕ_i and ϕ_j) lie close to each other. Equation (5.4-NET) expresses the homophilic assumption by considering that close ideologies enhance the probability of an existing connection :

$$Pr(y_{ij} = 1) = logit^{-1}\left(s_i + p_j - \gamma \left|\left|\phi_i - \phi_j\right|\right|^2\right). \tag{5.4-NET}$$

where $logit^{-1}(x) = [1 + \exp(-x)]^{-1}$ is the logistic function, and where the two variables p_j and s_i model respectively the effect of the political elites' popularity and the citizens' susceptibility to follow a Twitter account. The variable γ normalizes the effect of the ideological distance $||\phi_i - \phi_j||^2$.

Maximum likelihood estimation methods are usually intractable to derive accurate values for the different variables, given the large number of parameters involved and the shape of the likelihood function. Estimates $\hat{\phi}_i$ and $\hat{\phi}_j$ can however be obtained by averaging samples from the Bayesian posterior distribution [Barberá (2015)]. We derive network estimates $\hat{\phi}_{net}$ by averaging 800[11] generated samples from a Non-U-Turn sampler [Hoffman and Gelman (2014)]. It requires to determine priors and initial values for the model parameters. We consider normal priors on the model parameters as described in [Barberá (2015)]. The algorithm requires to provide initial values for the elites' ideologies. We propose initial values by performing a unidimensional correspondence analysis on the binary matrix Y. All the other parameters are randomly initialized as described by Barberá (2015).

Network data is more static—as social media users do not update their friends every day—than textual data, which can allow for more substantive and dynamic analysis of public opinion trends.

5.3.5 Towards a textual approach

Difficulties to handle online textual data come from the fact that they often include small and informal bits of texts. This is particularly the case for social media platforms like Twitter that limits users' posts to 140 characters. In addition to these limitations, social media users do not only post on political matter, making the messages even more difficult to analyze using existing textual analysis methods [Caton, Hall, and Weinhardt (2015)]. Of course, supervised textual models with high quality filtering can address some of these challenges by detecting a large proportion of non-political content [Lampos, Preotiuc-Pietro, and Cohn (2013); Toff and Kim (2013)]. However, supervised methods are most often based on static dictionaries that do not have the flexibility to deal with informal language, and might even introduce some of their authors' interpretation bias [Grimmer and Stewart (2013)]. Unsupervised methods are much better equipped to deal with such textual data.

[11] The samples are generated with 200 warm-up iterations followed by 800 iterations and a thinning of 2 on two different chains

The existing unsupervised content analysis methods require documents to have this politically dominant property. As we have seen, these methods are generally performed on large and formal documents like party manifestos [Slapin and Proksch (2008)]. Yet, most of the content posted on social media platforms are characterized by their their informality and are therefore not suitable to be analyzed by these methods. Therefore, even if the *Wordfish* algorithm has recently shown some promises in scaling ideological positions from messages posted by political elites on Twitter [Ceron (2017)], more needs to be done in order to extract reliable ideological positions from ordinary citizens' messages.

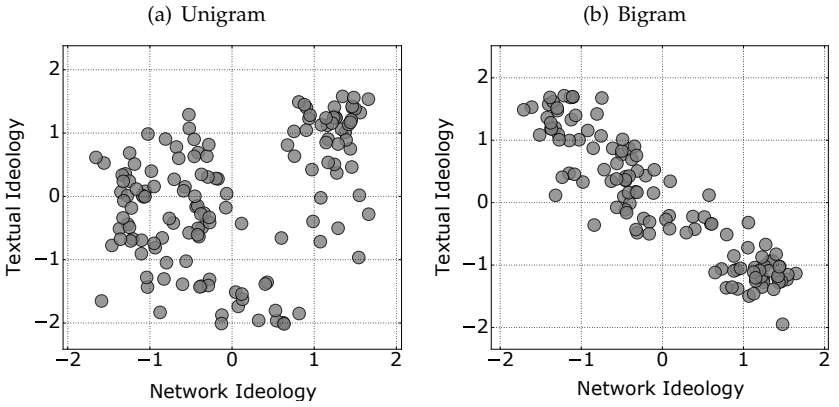


Fig. 5.4 Elites : illustrating the efficiency of N-gram term-document matrices. Comparison of network and textual based estimates of ideology for Canada 2015.

We attempt to generalize the existing textual approaches for social media posts by proposing two extensions. In a first time, we expand the content analysis by including groups of successive words instead of considering single words separately. In computational linguistics, a sequence of N adjacent words is referred to as an N-gram. Taking into account the co-occurrences with other words helps to distinguish semantically groupings [Brown, Desouza, Mercer et al. (1992)] and even to handle sarcasm better [Lukin and Walker (2013)]. Including N-grams substantially improves the quality of our estimates. As an example, Figure 5.4 compares network estimates to unigram ($N = 1$) and bigram ($N = 2$) textual estimates[12] for elites in the context of Canada 2015. We clearly observe that the textual estimates highly depend on the N-gram strategy. In this case, bigrams are strongly correlated with the network ideologies

[12]The generation of textual estimates from Twitter content is detailed in the next section.

whereas classical unigram methods fail to retrieve valid positions. Note that when the retweets are discarded from the original sample, we still observe an increase of correlation from $N = 1$ to $N = 2$, although the improvement is much more moderate.

Second, we propose to tackle the citizens' content by creating an unsupervised political dictionary from elites' tweets. In other words, we draw from the lexicon of political elites in order to build dynamic dictionaries from which we finally extract ordinary citizens' ideologies. Political expressions are varying over time, countries, and election contexts. It is therefore not suitable to work with one unique dictionary. A political dictionary is therefore created for each election. Ideally, generating new dictionaries should not require any human control, which allows to keep the process automatic.

5.3.6 Textual Estimates for Elites : inclusion of N-grams

We first record all the sequences of N adjacent words in the candidates' tweets for particular values of $N = 1$, 2 or 3. A series of preprocessing filters are then applied on the N-gram, as depicted by Figure 5.5. Only the dominant language is kept for analysis. Stop words are discarded and stemming is performed to improve the extraction accuracy. The remaining N-grams are then stored in lower case. During a second stage, we handle non-political terms by introducing a threshold parameter β. We discard N-grams that are not shared by at least β percent of the set of political elites. This step is called the β-filtering process. The process then builds a term-document matrix (Z^{el}) by matching the N-grams' occurrences to the corresponding politicians. We finally use the Poisson model described at (5.3-WORD) to generate the elites' estimates.

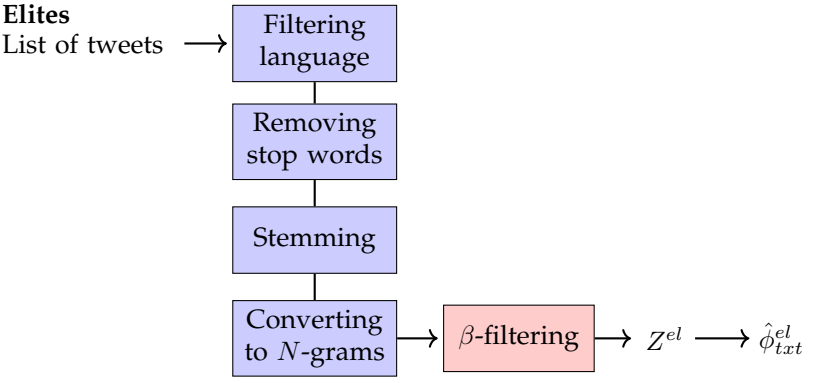


Fig. 5.5 Schematic overview of the textual method for elites

Figure 5.6 compares the performance of three N-gram sizes: unigrams ($N = 1$), bigrams ($N = 2$), and trigrams ($N = 3$). Each N-gram size is evaluated according to three criteria : *quality*, *classification improvement*, and *scope of information*. The case $N = 1$ corresponds to the classical approach designed for manifestos. In order to circumvent the comparison problem of multilingual textual analysis, the analysis is restricted to English-speaking members for New-Zealand 2014 and Canada 2015 and to French-speaking users for Quebec 2014.

The *quality* criterion measures the convergent validity of the textual estimates compared to other established ideological estimates (see Adcock, 2001). Ideally, we would compare our estimates to reference ideologies such as those derived from survey data (see Section 5.2.1) or from *roll-call* (see Section 5.2.2). Even if we can position political parties on issues using content analysis of their manifestos and other publicly available statements, the strict party discipline taking place in the contexts under study blurs individual candidates' positions by forcing them to be nearly identical to their related party. Therefore, to test the convergent validity of the textual estimates, we rely on the network estimates for politicians [Barberá (2015)]. We consider a social media extraction as effective if two elites with close reference ideologies also depict close estimates. The *quality* is then defined as the Pearson correlation coefficient between the vector of estimates derived from textual data from social media and the vector of estimates from network data. Its value ranges from 0% (poor quality) to 100% (high quality).

The *classification improvement*, denoted by C_{imp}, pictures the complementarity between textual and network estimates. It explains how well the textual estimates improve the detection of the party clusters compared to a simple network classification. This improvement is computed as follows:

$$C_{imp} = \frac{(a_{net+txt} - a_{net})}{(1 - a_{net})} \times 100\%$$

where $a_{net+txt}$ denotes the classification accuracy (see Section 5.4.1) obtained by using both network and textual estimates, while a_{net} denotes the accuracy by using only the network estimates. For the particular value $C_{imp} = 0\%$, no improvement is made by adding the textual features into the classifier. When $C_{imp} = 100\%$, the elites are perfectly classified ($a_{net+txt} = 1$) with both features
.

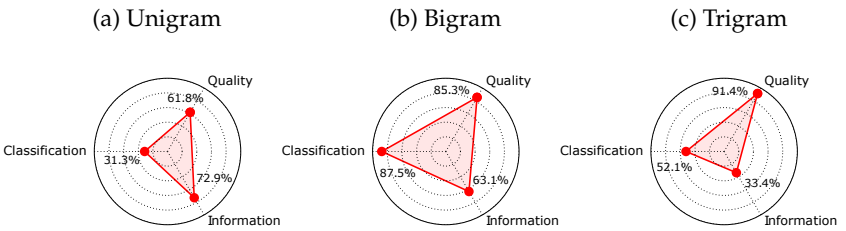


Fig. 5.6 Elites - Comparing the performance of the textual method for unigram, bigram and trigram strategies. Values are obtained by averaging over the three elections.

The *scope of information* records the percentage of the elites' sample[13] for which the textual method can provide an estimate. As the value of N increases (i.e., from unigrams to trigrams), the term-document matrix becomes sparser. This directly reduces the sample of elites for which an estimate can be provided.

Overall, Figure 5.6 shows that with the increase of N-grams, the *quality* also increases but this results in a substantive loss of information. In terms of classification, bigrams seem to perform better than single words and trigrams. Globally, working with bigrams faces the trade-off between validity and sample size. It outperforms other N-grams at classification while keeping decent *scope of information* and *quality*.

In addition, we look at the effects of the filtering conditions on the *quality* of the textual estimates. For bigrams ($N = 2$), low filters ($\beta < 3\%$) tends to keep noisy N-grams and higher filters ($\beta > 15\%$) are suppressing important information for the positioning. Consequently, the ideal dictionary should integrate this filtering trade-off. These effects are shown in Figure 5.7.

[13] That is, elites for which a network estimation is available.

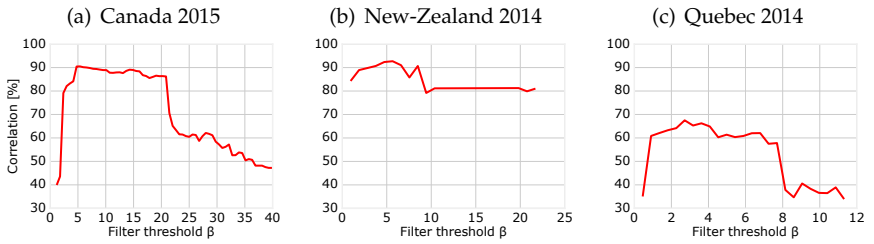


Fig. 5.7 Evolution of the quality of textual ideologies for increasing filtering conditions. The x-value represents the filter threshold β. The y-value displays the correlation between the network and textual estimates for the political elites.

5.3.7 Textual Estimates for Citizens : a political dictionary

The assumption of political dominance can be verified to some extent for political actors, but does not especially hold for ordinary citizens. As a result, it will be shown that simply building a term-document matrix from citizens' content does not lead to identify valid ideologies, which is due to the lack of political content. The problem is addressed by handling the citizens' textual data with past information computed on the elites, that is, the political dictionary. The dictionary consists of a list of elites' N-grams and their associated weights ($\hat{\theta}^{el}$) and popularity parameters ($\hat{p}^{el}$). The N-grams are those that were not discarded by the filtering process. Weights and popularities are directly computed from the Poisson model (5.3-WORD) by maximum likelihood.

The previous considerations lead us to create a dictionary for $N = 2$ and with $\beta = 5\%$. Table 5.2 shows a subsample of the political dictionary generated for Canada 2015. Samples of the New-Zealand 2014 and Quebec 2014 dictionary are available in Appendix C.1.

Bigrams on the left are those associated with the highest weights. They are all related to the NDP party. Stems like *leader - tom* and *canada - dream* are references to Thomas Mulcair – leader of the NDP party – and to his slogan *Building the country of our dreams*. The right part depicts bigrams associated to the lowest weights. Stems like *low - tax* or *conserv - protect* correspond to more conservative visions. Note that the dictionary creates two different items for bigrams composed of identical but inverted terms such as *ndp - cdnpoli* and *cdnpoli - ndp*.

In that sense, we propose to estimate the citizens' ideologies by taking into account the precomputed values ($\hat{\theta}^{el}$, $\hat{p}^{el}$) associated with each bigram present

Highest Weights				Lowest Weights			
Bigram		Weight	Popularity	Bigram		Weight	Popularity
ndp	cdnpoli	+6.44	-3.38	conserv	protect	-6.16	-2.67
leader	tom	+6.04	+5.42	tax	mean	-6.10	-3.03
cdnpoli	ndp	+4.46	-2.70	plan	protect	-5.74	-1.55
ndp	leader	+3.45	-2.27	low	tax	-5.58	-0.99
canada	dream	+3.31	-4.87	protect	econom	-5.41	-0.76

Table 5.2 Subsample of the political dictionary for Canada 2015. The table displays the 5 lowest and highest weighted bigrams in the stemmed form.

in the dictionary. The citizens' scaling process is illustrated at Figure 5.8. A similar preprocessing to the one developed for the elites is applied to the citizens' tweets. The building process of the citizens' term-document matrix Z^{ci} slightly differs than that for the elites. The matrix is built by matching the W terms present in the political dictionary to the corresponding citizens. As a reminder, the TDM entries z_{iw} count the number of instances of a specific term (identified by $w = 1, ..., W$) that appears in the aggregated content of citizen i.

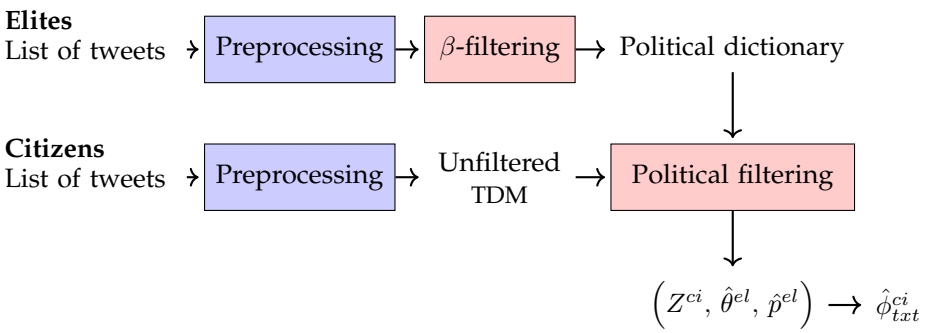


Fig. 5.8 Schematic overview of the textual method for the citizens.

The following scaling process estimates the textual ideologies from the TDM entries (z_{iw}) and from the precomputed values $\hat{p}^{el}$ and $\hat{\theta}^{el}$. We assume that the term occurence Z_{iw} follows a Poisson distribution:

$$P(Z_{iw} = z_{iw}) = \frac{\lambda_{iw}^{z_{iw}} \exp(-\lambda_{iw})}{z_{iw}!}.$$

where the event rates λ_{iw} are depending on the twitter user and on the underlying political term. As for model (5.3-WORD), we express the rates λ_{iw} as a functions of the citizens' ideologies, susceptibilities s_i[14], and on the precomputed popularities and term weights:

$$\lambda_{iw} = \exp(s_i + \hat{p}_w^{el} + \phi_i \hat{\theta}_w^{el}).$$

By considering the n citizens and the W political terms, the entire set of term occurences $Z_{(11)}, ..., Z_{(n,W)}$ has a joint frequency function expressed by the product of the marginal functions[15]. The log likelihood function l is therefore given by

$$l(\Lambda) = \sum_{i=1}^{n} \sum_{w=1}^{W} \left(z_{iw} \ln(\lambda_{iw}) - \lambda_{iw} - \ln(z_{iw}!) \right)$$

where Λ represents the matrix of term rates λ_{iw}.
The estimation of textual ideologies are obtained by maximizing the likelihood, which leads to the following optimization problem :

$$\underset{s_i, \phi_i}{\text{maximize}} \sum_{i=1}^{n} \sum_{w=1}^{W} \left(z_{iw} \, (s_i + \phi_i \, \hat{\theta}_w^{el}) + z_{iw} \hat{p}_w^{el} - \exp(\hat{p}_w^{el}) \, \exp(s_i + \phi_i \, \hat{\theta}_w^{el}) - \ln(z_{iw}!) \right)$$

where both terms $z_{iw}\hat{p}_w^{el}$ and $\ln(z_{iw}!)$ does not depend on the variables ϕ_i and s_i. Suppressing these terms does not alter the optimal solution and leads to the following convex optimization problem:

$$\underset{s_i, \phi_i}{\text{maximize}} \quad \sum_{i=1}^{n} \sum_{w=1}^{W} \left(z_{iw} \, (s_i + \phi_i \, \hat{\theta}_w^{el}) - \alpha_w \, \exp(s_i + \phi_i \, \hat{\theta}_w^{el}) \right) \quad \text{(5.5-CI-WORD)}$$

where the constant $\alpha_w = \exp(\hat{p}_w^{el})$ takes into account the N-gram popularity vectors estimated from elites' content.

We compare our estimations with the classical approach, in which we ignore the precomputed values and instead apply the Wordfish algorithm. Two different baseline strategy are proposed, depending on whether we consider the use of the political dictionary to create a term-document matrix (TDM). In the first strategy (*Wordfish-all*), the citizens' TDM is generated by keeping all the N-grams (i.e., following the same building process as for the elites term-

[14] We remind that the susceptibility is a factor that represent the publication activity of a user.
[15] Under the assumption of independence between the variables

document matrix Z^{el}). The second strategy (*Wordfish-political*) considers the adapted term-document matrix Z^{ci} built from the political dictionary.

As a point of comparison, we generate reference estimates from the election survey. In VPL's survey, the citizens express themselves on several political questions. The ideological estimates are built from citizens' responses to 30 survey questions on political issues. A undimensional factor analysis (5.1-PREF) is performed to extract the reference ideologies from the recorded answers. We use an expectation-maximization algorithm to solve the maximum likelihood problem that generates these estimates [Barber (2012)]. We measure the quality of textual estimates as the Pearson correlation coefficient between the vector of textual estimates and the vector of reference ideologies.
Table 5.3 shows the textual method's performance (5.5-CI-WORD) compared with the two baseline strategies. The textual extraction clearly outperforms both baselines, especially for Quebec 2014. Nevertheless, we observe only moderate correlations with the reference ideologies. We also add to our analysis the network estimates as generated by Barberá (2015), which are strongly correlated with the reference ideology ($\rho > 0.6$). This result suggests that by looking solely at one's public network of friends we are able to derive accurate positions.

Method	Canada 2015	New-Zealand 2014	Quebec 2014
Textual	44%	40%	20%
- Wordfish-all	30%	30%	8%
- Wordfish-political	39%	39%	9%
Network	63%	77%	76%

Table 5.3 Citizens - Quality of the network and textual scaling methods for each election. Results are expressed for a bigram dictionary. Percentages indicate the Pearson correlations (p-value < 0.05) between the ideologies extracted from Twitter data (posts and network) and the reference ideologies based on survey data.

One may wonder if a better measure of ideology can be obtained by combining network and textual positions into one single estimate. A classical way of merging estimators is to consider the set of estimators $\hat{\phi}_\lambda$ generated by convex combinations [Struppeck (2014)]. This family of ideological estimators is described by:

$$\hat{\phi}_\lambda = \lambda\,\hat{\phi}_{net} + (1-\lambda)\,\hat{\phi}_{txt} \quad , \quad \lambda \in [0,1] \tag{5.1}$$

.

We compare the efficiency of combining ideologies by analyzing the correlation of the new estimates $\hat{\phi}_\lambda$ with the reference ideologies according to multiple values of λ (Figure 5.9). The optimal *quality* for Canada 2015, New-Zealand 2014 and Quebec 2014 is respectively reached for $\lambda = 0.86$, $\lambda = 0.87$ and $\lambda = 1$. None of these combinations leads to enhance the network performance by more than 1%. This suggests that combining estimates does not lead to a significant improvement.

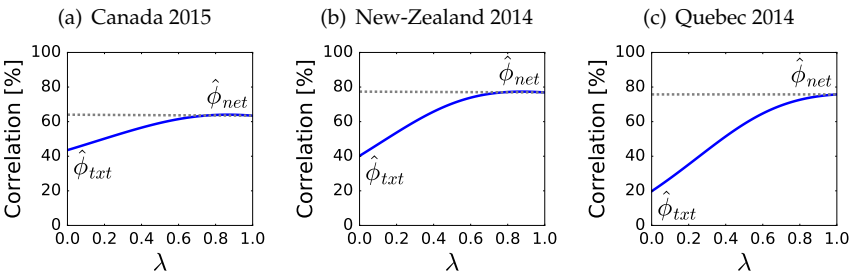


Fig. 5.9 Citizens - Assessing linear combinations of textual and network ideologies. The x-axis displays the parameter λ in equation (5.1). The textual ideology corresponds to the case $\lambda = 0$. The network ideology corresponds to the case $\lambda = 1$. The y-axis displays Pearson correlations (p-value < 0.05) between the linear combination and the reference ideologies. The dashed line depicts the correlation for $\lambda = 1$.

We provide two alternative scenarios to explain the moderate correlations observed in the textual case. Either the presence of non-political bigrams generates noisy estimates that explain the lower correlations observed for the textual estimates. Or the derived positions can still contain relevant information about individuals, but do not especially capture the same ideological dimension as the one captured by the reference ideology. Therefore, the combination of these two types of information could be complementary. This means they could lead to better performance when we try to predict phenomena related to political ideologies such as voting behavior, as we will point out later.

5.3.8 Interpreting the political landscape

We end this section by visually displaying the political landscapes obtained from social media data and by discussing the positions of the party clusters. The analysis first focuses on the English-speaking elections (New-Zealand 2014 and Canada 2015), for which the elites are represented in Figure 5.10. Links to interactive graphs are provided through QR-codes for the convenience of the reader. They redirect to a webpage allowing the reader to display the elites' identity when *mouseovering* each estimated point [16]. Hyperlinks are also available at Appendix D.

It is apparent that the point estimates – showing each candidate's position measured using two fundamentally different scaling methods – are strongly correlated. This emphasizes a linear relationship between the estimates and indicates that they seem to capture a similar ideological dimension, which confirms implicitly the validity of both methods.

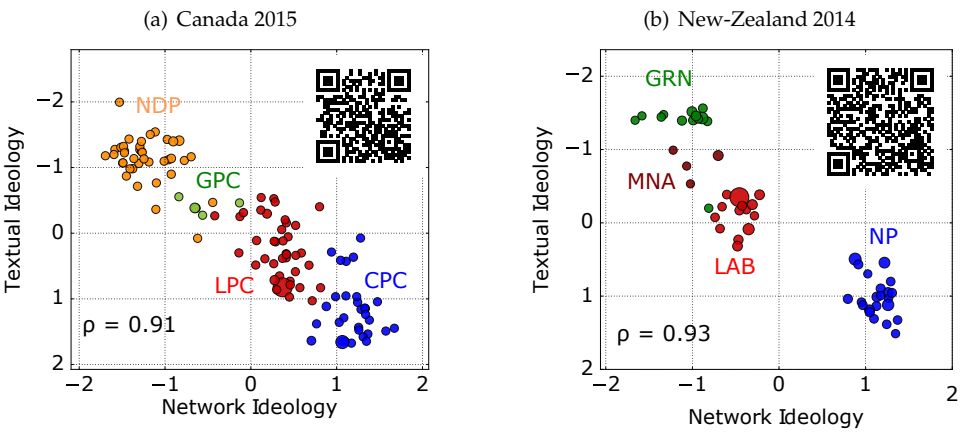
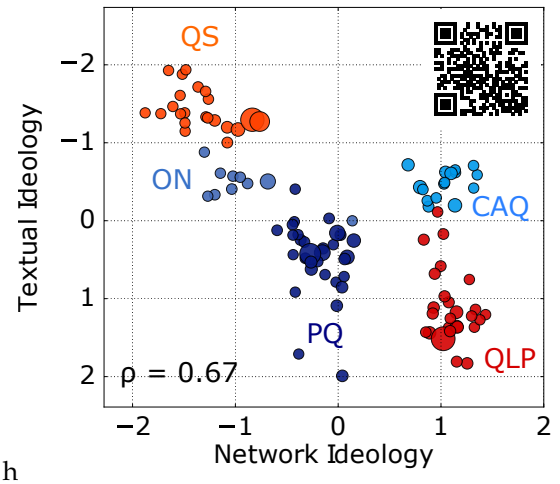


Fig. 5.10 Political elites - Comparison of network and textual based estimates of ideology. The x-axis shows the standardized position of political elites on the unidimensional latent space derived from network data. The y-axis shows the standardized position of political elites on the unidimensional latent space derived from textual data. The dot sizes are proportional to the number of followers. The upper right corner features a QR-code leading to an interactive version of the graph. The lower left corner displays the Pearson correlations (p-value < 0.05) between the estimates.

By taking a closer look at the different political contexts under study, we also

[16] Names are displayed by clicking on the dots when using a smartphone. The absence of some leaders is explained by the fact that some politician's account did not pass the filtering criteria.



h

Fig. 5.11 The particular case of Quebec 2014. The x-axis shows the standardized position of political elites on the unidimensional latent space derived from network data. The y-axis shows the standardized position of political elites on the unidimensional latent space derived from textual data. The dot sizes are proportional to the number of followers. The upper right corner features a QR-code leading to an interactive version of the graph. The lower left corner displays the Pearson correlations (p-value < 0.05) between the estimates.

see that these methods can identify clusters of points belonging to the same political party, even though the within-party correlations are weak. These are not only consistent with what is expected by conventional wisdom on party positions, but also with estimates derived from party positions towards political issues based on the survey conducted by VPL. In both context, when we compare the mean for each party's cluster on Figure 5.10 with the policy positions in Figures 5.1 and 5.2, we can see that these points depict the ideological landscape that is expected. We observe, however, an inverted position for the Green and NDP party in the context of Canada 2015. This inversion is partially explained by the fact that the Green's party leader – Elizabeth May – has recently given to the party a more centrist position [Whitehorn (2015)]. Both methods are able to capture a valid ideological landscape from free publicly available social media data.

We find a slightly divergent result in the 2014 Quebec context. Contrary to the previous two political contexts, the network- and textual-based methods reveal different ideological landscapes in this case. For instance, the network estimates (x-axis) taken solely cannot differentiate the QLP from the CAQ (as

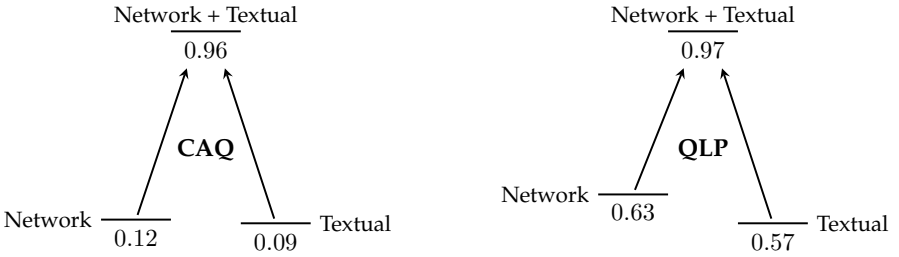


Fig. 5.12 Quebec 2014 - Complementarity of network and textual estimates to detect the CAQ and QLP political clusters. Depicted values corresponds to the average classification efficiency measured by the AUC (described at Section 5.4.1)

observed in Figure 5.11). This result is interesting as these two parties are known to share similar positions on socio-economic issues. Similarly, although the x-axis clearly separates the QLP from the CAQ, the latter party becomes indistinguishable from the ON when we only consider the textual positions.

These results might suggest that these two methods, when taken together, can complement each other when capturing the ideological landscape. This is shown in Figure 5.12: the network estimates as well as the textual estimates display low classification performances[17] when the features are taken apart. However, coupling the features allows to reach an interesting precision ($> 95\%$ for CAQ and QLP candidates).

That being said, the two methods identify relative clusters that are consistent with the parties' positions as expected by conventional wisdom and as depicted by the survey-based method. And that, in two different languages: French and English. The results suggest that a unique dimension might not be sufficient to capture complex ideological landscapes.

5.4 Application to voting behavior

Predicting voting behavior remains a challenging empirical task. One of the reasons that explains scholars' interest in ideologies lies behind their ability to predict voting behavior [Downs (1957); Jessee (2009)]. Any single validation procedure is also not sufficient to establish the validity of a measurement. To further this validation process, we assess the nomological validity of our

[17] Classification performance is measured in terms of AUC, further detailed in Section Classifying elites and citizens.

estimates by testing their power to predict individual-level voting behavior. Contrary to the convergent validity, which looks at the correlations between multiple indicators of the same concept, the nomological validity compares different phenomena that are linked together by an explanatory relation [Adcock (2001)].

5.4.1 Classifying elites and citizens

We describe a machine learning classification task to investigate the ability of the ideological estimates to correctly classify party affiliation (for elites) and predict vote intentions (for citizens). The machine learning approach is preferred to a descriptive statistical study since we focus on predicting an outcome rather than providing an explanatory model.

The textual and network ideologies—as well as the survey ideologies for the citizens—are now seen as features or predictors while the parties can be seen as categories or labels. Predicting affiliation and intentions can be formulated as a multi-label classification task since more than two parties are generally involved. We deal with the multiple labels by introducing successive and independent binary classifiers [Read, Pfahringer, Holmes et al. (2011)]. Each party is associated with one single classifier. In the case of citizens, the classifier predicts for every citizen whether they intend to vote for an underlying party. For political elites, the decision boundary attempts to separate the candidates belonging to a specific party from the rest of the other elites. The classification performance is sometimes assessed by computing the *accuracy* of the classifier. The accuracy is defined as the fraction of all instances that are correctly categorized. However, predicting with a high accuracy does not prevent from generating a large amount of false negatives and false positives. The two conventional indicators that allow to detect these high rates are the *precision* and the *recall* of the classifier. To illustrate, we assume that the classifier related to a party A outputs a subset of citizens (noted $\tilde{V}_A$) for which it predicts a vote intention for A. Among the set $\tilde{V}_A$, the *precision* measures the fraction of citizens that are indeed expressing a vote intention for A. By contrast, the recall measures among the A-voters the fraction for which the prediction was correct (i.e., the fraction of A-voters belonging to the set $\tilde{V}_A$).

The analysis only includes members who reported one of the major parties in their voting intentions. Also, training sets with highly unbalanced class sizes generate important biases [Huang and Du (2005)]. The classification analysis is therefore performed on parties with at least 20 voters. This restricts

the classification analysis to 8 major parties for the three electoral contexts[18] and sample sizes of $n = 796$ (Quebec 2014), $n = 123$ (New-Zealand 2014) and $n = 1717$.

The use of *Support-Vector Machine* classifiers with Gaussian kernels is recommended to handle a smaller number of features and medium size training samples [Hsu, Chang, Lin et al. (2003)]. We prevent over-fitting by following a threefold cross-validation process to train the SVM classifier. That is, the training of the classifier is performed on a dataset including $2/3$ of the sample (training set) and the prediction performance is assessed on the remaining accounts (validation set). The performance of one feature is assessed by micro-averaging the performance indicators assessed on the validation set.[19] Our classifier requires to determine two parameters : the variable κ composing the Gaussian kernel $K(\phi_i, \phi_j) = \exp(-\kappa(||\phi_i - \phi_j||^2)$, and the value of C which controls the cost of misclassified training samples [Smola and Vishwanathan (2008)]. We use the values of $\kappa = 1/2$ and $C = 1$ in the present analysis. Using an SVM classifier implies also to determine a threshold value on the decision boundary. Applying different thresholds generates multiple couples of precision and recall, leading to a precision-recall curve for each party.

The classification efficiency of a particular feature is assessed by micro-averaging the precision-recall curves for each election, which is illustrated in Figure 5.13. These are compared to two classical baselines: the average precision corresponding to the random classification of the members (gray area)—which corresponds to 1 over the number of parties—and the precision obtained by training the SVM on the label distribution (the gray dotted line). The average performance of a feature is assessed by computing the *Are Under the Curve* (AUC), obtained by integrating the precision function over the recall interval. We observe that the *network* and *survey* based estimates show similar performances, from intermediate precision (AUC > 0.65 for Canada) to high precision (AUC > 0.80 for Québec and New Zealand). These two features outperform the *text-based* approach, which reflects little if any improvement compared to the label-trained baseline.

5.4.2 Predictive Power and Complementarity

As illustrated in the previous sections, new data sources combined with novel scaling methods allow the estimation of ideological positions that can predict votes without the cost of designing and administrating complex surveys. Fur-

[18] NP,PQ,NDP,LAB,PLQ,CPC,LPC and CAQ

[19] Micro-averaging is performed by averaging the curves in proportion to the number of voters for each party.

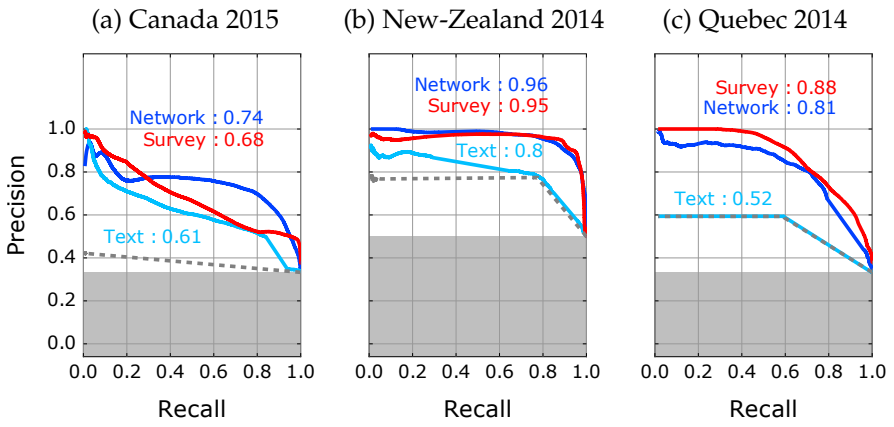


Fig. 5.13 Citizens - Average efficiency of citizens' ideologies to predict vote intention. Each curve displays the micro-averaged precision and recall obtained by varying the threshold of a Support Vector Machine classifier with a Gaussian kernel. The classifier is trained using 2/3 of the citizens' ideologies and the precision-recall couple is computed on a 1/3 validation set. The gray area serves as a lower bound and is derived from a random classifier. The gray dotted line represents a baseline when the classifier is trained on the labels' distribution. Values above the graph represent the area under each curve (AUC).

thermore, the different sources of information available can complement each other. In order to investigate the complementarity of these methods, we further evaluate the predictive value for vote choice for each of the two methods individually, and when combined.

The Venn diagrams (Figure 5.14) illustrate the average prediction efficiency based on network, textual, and survey ideology estimates. This allows us to show the quality of the predictions for each data type taken individually and for any combination between these. The metrics displayed are an evaluation of the quality of the predictions based on the AUC of precision and recall curves within each context. These metrics evaluate the proportion of time the trained algorithm guesses the voting intentions of an individual correctly. The advantage of this metric is that it is less affected by sample balance than simply using the prediction accuracy. As described previously, the crossvalidation process handles over-fitting so that the results of the combinations of these estimates are not trivial and actually results from the additional information present in the feature.

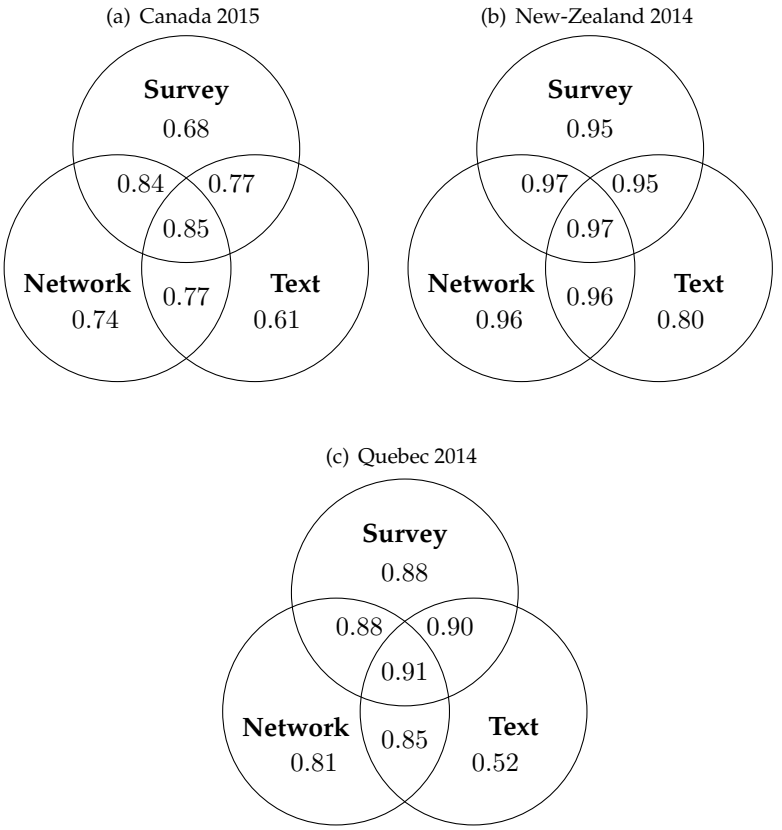


Fig. 5.14 Venn diagrams illustrating the complementarity between the ideology estimates to predicting voting intentions of citizens. The metric value inside each set represents the prediction efficiency measured by the area under the curve (AUC).

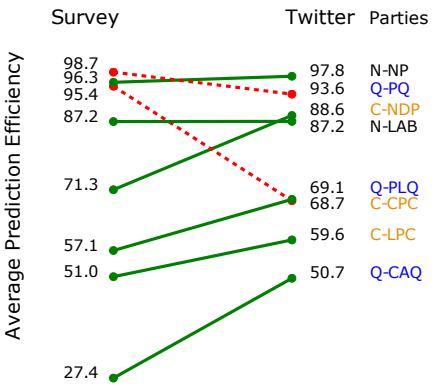


Fig. 5.15 Comparison at the party level of citizens' Twitter and Survey prediction efficiency. The values displayed correspond to the area under the curve of the precision and recall curves related to each party.

When we look at the AUC for each type of data on Figure 5.14, we can see that the predictions based on network ideology estimates, taken solely, can outperform the predictions based on survey ideology estimates. This is the case for Canada in 2015 and for New Zealand in 2014. The 2014 Quebec network- or textual-based predictions do not outperform the survey estimates. Furthermore, we see that the quality of the predictions improves when combining the estimates . The only exception would be for New Zealand where it is clear that the text estimates do not add any information to network nor to survey estimates. In Quebec, the text estimates add information to the network and survey estimates. This result is consistent to what is shown for elites on Figure 5.10. This shows the positive effect of combining different types of data that are available on social media. That is, the quality of our predictions improves when combining network and text estimates.

To further examine the benefits of combining social media data types, Figure 5.15 illustrates the performance of our classifier at the party-level. The values displayed correspond to the AUC of precision and recall curves related to each party. The plain lines highlight the parties where Twitter estimates[20] perform better than the conventional survey estimate. Conversely, the dashed lines indicate when Twitter information leads to higher prediction errors. We

[20]That is, textual and network estimates.

can see on Figure 5.15 that Twitter-based predictions of voting intentions generally outperform survey predictions. Even though the evidence does not allow us to generalize this pattern, it is worth noting that the dashed lines are related to the PQ and the CPC, two incumbent parties that were the least active on social media relative to other parties. There is also a noticeable higher efficiency rate for smaller parties (CAQ, NDP). This could partly be explained by the fact that they have a more important online presence, in terms of quantity of tweets published by their respective political elites. Parties with smaller campaign budgets will tend to rely more on cheaper social media strategies than richer parties [Haynes and Pitts (2009)]. Therefore, they tend to publish larger quantities of information on social media in order to try to reach a broader audience at lower costs. As we have more data for these smaller parties, this can help the classifiers to be more efficient at identifying individuals with intentions to vote for such parties. The more information we have to train the classifiers for these parties translates in less classification errors. This is not the case with survey data as everyone answers questions related to specific policy issues relevant to the political campaign.

5.5 Elections without survey data : the Belgian context

This last section focuses on the scaling of the Belgian parliamentarians' and ministers' Twitter accounts. The particularity of this analysis lies in the fact that the scaling, performed on elites, is carried out outside an election period and without access to survey information.

The textual extraction is applied on a series of tweets published through a 3-month period, between January 1, 2017 and April 1, 2017. The network algorithm takes as input a binary matrix resulting from a set of citizens, each one of them following at least two political actors. Unlike the study performed on the Vox Pop Labs' survey, this time we do not have access to a list of respondents. The binary matrix is therefore obtained by aggregating the whole set of followers[21] of each politician. Followers that are not geolocated in Belgium are discarded, as well as accounts that do not satisfy the activity constraints suggested in Section 5.3.3. The major political parties in Belgium are either French-speaking or Dutch-speaking. The study is performed separately for each community (Wallonia and Flanders).

The French speaking parties Figure 5.16 (a) represents the network and textual estimates for the six major parties in Wallonia. Although a weak correlation

[21] Followers are obtained for each elite using the Twitter API.

between the estimates is discerned, the party clusters are clearly identifiable. The Socialist Party (*PS*, in red) is a center-left party. Socialist politicians display a wide variation of network estimates, which could indicate a broad-based electorate. By contrast, the center-left Reformist Movement (*MR*, in blue) is characterized by a smaller variance in its electorate, which is compensated by a larger spectrum on the textual axis. The Humanist Democratic Centre (*CDH*, in orange) and Independant Federalist Democratic (*Défi*, in pink) share the midpoint of the political space. The Walloon Green party (*Ecolo*, in green) stands out from the other parties at the textual level. The Workers' Party of Belgium (*PTB*, in brown) occupies a far-left position that is perceptible on the network scale, though with a high variance amongst the political elites.

Some politicians are located further away from their party kernel. As an example, *Elio Di Rupo* and *Paul Magnette* (the two red right upper points) are seen on the network scale as outliers to the PS cluster. While politicians such as *Laurette Onkelinx* and *Jean-Claude Marcourt* are seen as the voice of the Socialist party (and accordingly stand a central location in their cluster), *Di Rupo* and *Magnette* are perceived as intellectuals who can afford to take a slightly different stance[22].

The Dutch speaking parties The six Flemish political parties analyzed in Figure 5.16 (b) are the New Flemish Alliance (*NVA*, in yellow), the Vlaams Belang (*VB*, in black), the Open Flemish Liberals and Democrats (*Open VLD*, in blue), the Christian Democratic and Flemish (*CD&V*, in blue), the Flemish Green party (*Groen*, in green) and the Socialist Party Differently (*sp.a*, in red). The correlation displayed ($\rho = 0.84$) is much stronger than for the Walloon elites.
Surprisingly, the right party VB is located on the left of the center-right party NVA. This is partially explained by the fact that the party recently started to mitigate their speeches for a de-demonization of their extreme position, as it is the case for the *Front National* in France. Scholars have observed these last few years that the NVA strategy slightly differs from those of the other parties. In particular, the NVA appropriates some of the CD&V's and the VB's votes by playing on a large number of issues, which translates into a high variation on the y-axis. On the other side of the map, the other parties highly overlap each other. This is particularly conspicuous for the CD&V and the Open VLD on the network axis.

To summarize, the analysis performed on the Belgian case highlights the party

[22]See for instance the analysis made in https://goo.gl/v6U6ix by Gaël Brustier. Website last visited on 23 July 2017.

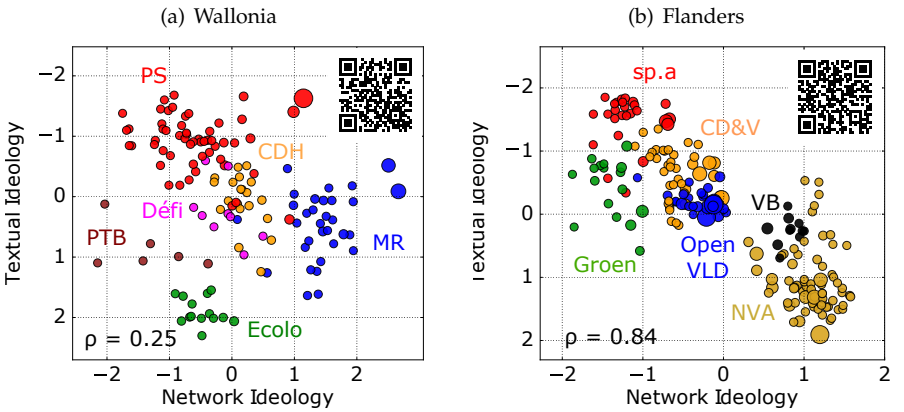


Fig. 5.16 Belgium - Comparison of network and textual based estimates of ideology. The x-axis shows the standardized position of political elites on the unidimensional latent space derived from network data. The y-axis shows the standardized position of political elites on the unidimensional latent space derived from textual data. The dot sizes are proportional to the number of followers. The upper right corner features a QR-code leading to an interactive version of the graph. The lower left corner displays the Pearson correlations (p-value < 0.05) between the estimates.

clusters but displays smaller correlations than for the New-Zealand and Canadian case, especially for Wallonia. One would argue that Twitter content outside an election context is much less polarized, resulting in a decrease of the textual estimates' quality. From the network point of view, there are actually no guarantees that the aggregated set of followers do not include a significant proportion of fake users (i.e., Twitter bots), by contrast to the survey respondents of the three previous election contexts.

5.6 Concluding remarks

In this chapter, we introduce and evaluate a method to ideologically scale non-politically dominant content such as messages posted on social media by citizens. This method can estimate reliable and valid ideological positions for different types of political actors in different contexts and languages. Within each context, the resulting estimates are both correlated with other conventional survey measures of ideology (convergent validity) and predictive of vote intentions (nomological validity). These multiple validation procedures provide further evidence on the validity of these methods to extract ideological positions. Bigrams show on average the best performances in terms of *quality*

of estimates, *classification improvement*, and *scope of information*, when applied on sparse content such as Twitter posts. In addition to the content posted by individuals on social media, we also illustrate the complementary of other sources of information also available on social media, such as network information. The combination of these two sources of information in a single model, generally outperforms the reference survey-based ideology when predicting vote intentions. While network estimates have a higher predictive power, the textual estimates have the potential of allowing for real time analyzes of trends in the public opinion.

We believe that the new method presented in this chapter has the potential to study a wide range of political actors on which we might not have reliable measures of ideology. Its scalability and multilingual capacity also suggest that the method could be used to study political contexts where it is very difficult to conduct conventional surveys. Finally, we can expect future research to expand classifiers—that were previously trained on several past and known contexts—to provide accurate individual voting intentions for upcoming elections. These new publicly available sources of information, when used with proper methods, have the potential to enable the study of public opinion at the individual level, in real time, and thus, fundamentally enhance our understandings of public opinion dynamics.

Chapter 6

Memory in inter-communication times

One of the popular research topics on networked humanity is to understand how people interact and communicate [Blondel, Decuyper, and Krings (2015)]. The importance of understanding communication patterns has already been recognized as a proxy for inferring information on the user, for example, on their social group [Blumenstock, Gillick, and Eagle (2010)] or gender [Aledavood, López, Roberts et al. (2015); Kovanen, Kaski, Kertész et al. (2013)]. Similarly, analyzing airtime credit purchase patterns lead to derive social indicators that would otherwise be extremely costly to find using classical methods based on census [Decuyper, Rutherford, Wadhwa et al. (2014)]. Communication patterns are not only important for inference and for extracting information. It has been shown that they have a very strong effect on the actual dynamics of the network. On the one hand, they determine the way information [Karsai, Kivelä, Pan et al. (2011)] or diseases [Rocha, Decuyper, and Blondel (2013); Salathé, Kazandjieva, Lee et al. (2010)] spread over the network. Moreover, the robustness of network connectivity has been found to be reinforced thanks to the waiting time distribution observed [Takaguchi, Sato, Yano et al. (2012)].

The experiments and models designed in the previous chapters did not incorporate exhaustively the time structure present in human communication processes. During the crowdsourcing experiment, the interactions between the players are simplified when compared with how people behave on real social networks. At each round, the expressed opinions are all displayed at the same time for the whole group instead of appearing sequentially for the participants. Analogously, the textual extraction presented in Chapter 5 is

based on a term-document matrix that does not take into account the tweets' inter-arrival time. Hence, we investigate in this last chapter the modeling of inter-event times based on various commenting activities observed on social media. We propose a stochastic model that incorporates memory for posting events generated on these platforms. Starting from exploratory observations on the series of waiting times, we provide an intuitive explanation in terms of a simple two-state Markov chain process. We finally explain the observed waiting times as generated by a time homogeneous Markov Chain with continuous state space.

6.1 Burstiness in human communication

One way to represent a series of events is to take a snapshot of the communication within a time window and evaluate the events that occurred in a static way, as done previously when building term-document matrices. Alternatively, we may view the communications as a process evolving in time. For example, in the case of a homogeneous Poisson process where events happen uniformly and independently over a time period, we can describe the sequence of events as a process evolving in time with independent exponential waiting times. For human communication, it is not considered as a homogeneous process, instead, a bursting behavior has been observed. This has been translated as independent waiting times following a power-law distribution [Johansen (2004)], which roughly means that the density of the waiting time τ decreases as $\tau^{-\gamma}$ for some $\gamma > 1$. To illustrate, series of events generated by a Poisson process are compared in Figure 6.1 to power-law distributed events. The heavy-tailed distribution (b) is mainly characterized by events occuring in bursts and displays very long delays when compared to the Poisson process.

Scholars have looked into intuitive models that may explain the heavy-tailed distribution observed for human activity patterns. In an early work, Barabási (2005) draws a parallel between human decisions and queueing theory in order to explain the observed inter-event time distribution. The paper describes human activities as a list of tasks associated with different priority levels. By discussing the variability of the task queue length, the author assumes the existence of two universality classes, corresponding either to a power-law with coefficient $\gamma \approx 1$ (fixed queue length) or $\gamma \approx 1.5$ (variable queue length). Other studies tend to provide alternative explanations to the heavy-tailed distribution for human activities that does not fall into the category of the task-based approach [Ming-Sheng, Guan-Xiong, Shuang-Xing et al. (2010); Zhou, Kiet, Kim et al. (2008)]. They propose interest-driven models which depict how the activity level is related to the power-law exponent γ.

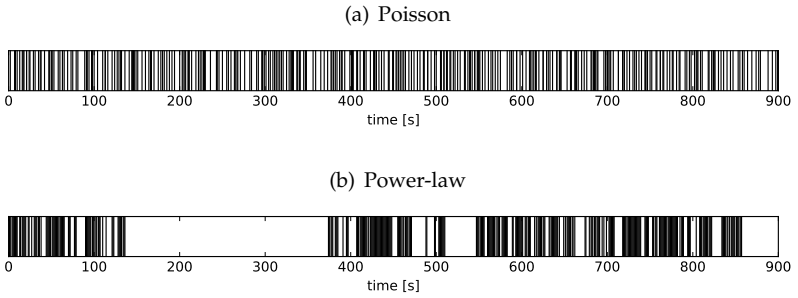


Fig. 6.1 Sequences of events generated by either a Poisson process (a) or by a power-law distribution (b). Both figures display 350 events (vertical lines). The Poisson waiting times are obtained with parameter $\lambda = 2.5$. The power-law waiting times are obtained with $\gamma = 2.5$ on the truncated interval $[1, m]$ with $m = 14$ hours.

Research has been done on the various features having effect on the waiting times. It is clear that human activities are subject to the effect of circadian and weekly cycles, which can be integrated in cascading Poisson processes [Malmgren, Stouffer, Motter et al. (2008)]. There is a memory effect present in human dynamics, which is, however, much more subtle than the inherent strong memory in natural phenomena [Goh and Barabási (2008)]. Nonetheless, it has been shown that people modify their activity rate based on perceived information concerning their past activity patterns [Vazquez (2007)]. The queueing theory analogy [Barabási (2005)] also introduces an implicit dependence for processes structured as the arrival and completion of tasks.

However, these approaches have two major drawbacks. First, when investigating online activities, we observe a clear dependence between pairs of consecutive activities. The current models do not incorporate those dependencies. For instance, Vazquez (2007) considers a long memory of an initial state but disregards the effects of later events. The second drawback results from a more theoretical consideration about the use of power-law distributions in the modeling of inter-event times. In principle, a power-law distribution can be used only on an interval bounded away from 0. Therefore when fitting the waiting times a cutting parameter has to be chosen. A high-cutting value discards useful data samples while a lower value leads to a biased estimation of the power-law exponent [Clauset, Shalizi, and Newman (2009)]. This already shows that there is a need to handle separately the shorter and longer inter-event times.

6.2 Temporal patterns in online activities

We focus our work on two popular social media: the social network *Twitter* and the news aggregator *Reddit*. We gathered data in a 6-month period, between April 1, 2016 and September 30, 2016. Clearly the night period is misleading, as it gives an extreme long waiting time corresponding to sleeping. Therefore, for each user, we cut the series of events into daily blocks and treat them separately. In each block, we also discard events generated outside the range $8h - 22h$. This could introduce a slight bias towards shorter times but is a simple and robust way of filtering out the circadian rhythm. Only users who have at least 1000 activities during the data collection period are kept for the analysis. We should acknowledge that this filtering condition substantially reduces the sample size and targets only strongly active members. It is nonetheless required to guarantee enough waiting samples per user.

We finally compute the elapsed time between all subsequent pairs of events for each "user/day" couple. This gives us multiple sequences of inter-event times $(T_1,\ T_2, ...,\ T_n)$. As no event occurs at the same time and since the timestamps are measured in seconds, all inter-event times verify $T_i \geq 1$.

Twitter posts The social network Twitter allows members to post short messages to the whole community. Each post is limited to 140 character content, which makes Twitter a fast interface for transferring short snippets of information. The website provides free access to the activity timestamps through the Twitter REST API. We extracted a total of $n = 4796$ geotagged and highly active users, tweeting from 9 different countries[1]. These users have been generated by aggregating the followers of the 100 most followed accounts for each country[2]. Due to the rate limits imposed by the REST API, we were only able to handle a small random sample ($\simeq 1\%$) of the total number of accounts ($> 10^8$), from which we discarded users without geolocalisation or generating less than 1000 tweets or retweets in the period of study.

Reddit comments Reddit is a website that allows members to submit links or textual content to the community that in turn react with public votes or

[1]The 9 countries are : France, Spain, Belgium, Italy, Germany, Portugal, United Kingdom, Netherlands and Canada

[2]These accounts were obtained from http://twittercounter.com/ on October 10, 2016

comments. One of the major differences with Twitter is that Reddit is subdivided in specific categories, called subreddits. This makes Reddit a news and information aggregator.
The Reddit events are recovered from a publicly available database regrouping about 1.7 billion comments published from January 2005 to December 2016. All comments and associated timestamps are accessible through the Google webservice *BigQuery*[3]. In total, the analysis counts $n = 3081$ Reddit users which passed the 1000 comments filtering restriction.

We have seen previously that opinions are often distributed around a truth value far from which they rarely lie. This fact contributes to the well-known wisdom of the crowd effect. However, not all phenomena are following this pattern. As far as communication activities are concerned, a heavy-tailed distribution as displayed in Figure 6.2 might be observed.

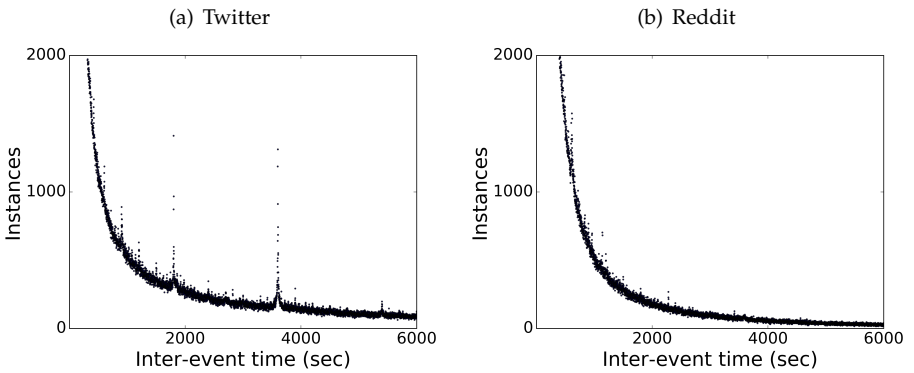


Fig. 6.2 The heavy-tailed distribution of inter-event times for the Twitter and Reddit datasets. The Twitter plot displays a noisy behavior at the values $T_i = 1800$ sec $(30$ min$)$ and $T_i = 3600$ sec $(1$ h$)$, which is suspected to be generated by fake users (bots).

Formally, we say that a variable T_i follows a power-law if it is drawn from a distribution that can be expressed as:

$$\rho(t) \propto t^{-\gamma}.$$

where one usually dictate $\gamma > 1$ for the power-law *exponent*. The probability density function is defined on an interval bounded away from 0, that is, for

[3]Tables are available at https://bigquery.cloud.google.com/table/fh-bigquery:reddit_comments.2015_05

values bigger than some positive t_{min}. These last two conditions allow to compute a normalizing constant :

$$\rho(t) = \frac{\gamma - 1}{t_{min}} \left(\frac{t}{t_{min}} \right)^{-\gamma}. \tag{6.1-POW}$$

One can easily explore power-law behavior by plotting the data samples on a doubly logarithmic scale. In other terms, taking the logarithm on both sides of equation 6.1-POW highlights a linear relationship $\ln \rho(t) = -\gamma \ln t$, up to a constant k. This linear relation is perceptible on Figure 6.3 for the Twitter and Reddit events.

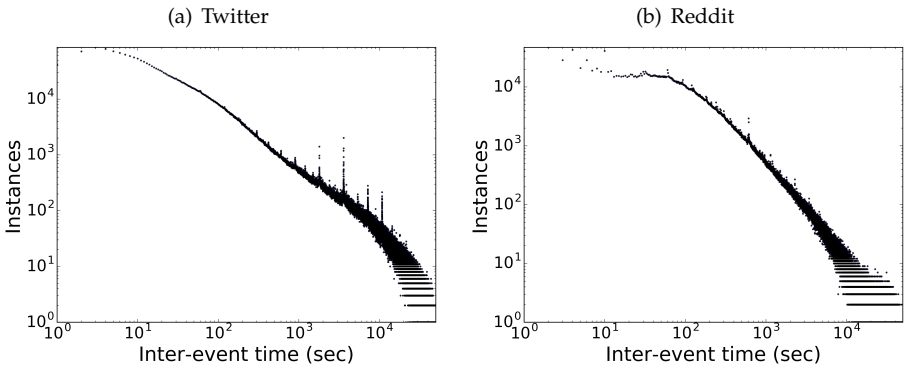


Fig. 6.3 The linear relation of inter-event times on a doubly logarithmic scale.

An effective way to infer γ from empirical data is to make use of maximum likelihood methods [Clauset, Shalizi, and Newman (2009)]. The ML estimator on a data sample of n waiting times $(t_1, ..., t_n)$ is given by :

$$\gamma = 1 + n \left(\sum_{i=1}^{n} \ln \left(\frac{t_i}{t_{min}} \right) \right).$$

Determining the value of t_{min} requires to make a trade-off between biasing the estimation of γ and discarding legitimate data samples. This issue is counteracted in the new approach that we suggest.

6.3 Empirical observations of time dependence

We analyze the time dependence by defining a variable – the threshold time t_{thres} – that splits the waiting times into 2 categories. The *short* waiting times

are those who satisfy the inequality $T_i < t_{thres}$, whereas the *long* waiting times correspond to the case $T_i \geq t_{thres}$.

We question the independence assumption of waiting times for the following reason. When looking at the sequences of waiting times, we observe that short waiting times are usually more likely to be followed by other short waiting times than what is predicted by an independent process. We formalize this by fixing arbitrary threshold values and by counting the number of occurrences of two subsequent short waiting times. One compares these occurrences with what the standard independent model would suggest. That is, we consider the ratio

$$r = \frac{p_{SS}}{p_S \cdot p_S}$$

where p_S represents the probability of generating a short waiting time and p_{SS} the probability of generating two consecutive short waiting times. For independent waiting times, the ratio r would be exactly 1 in theory and near 1 statistically. The resulting histogram (Figure 6.4) displays ratios well above 1 for most users. This is verified for both datasets and for arbitrary threshold values, which reinforces our intuition concerning the time dependence. Note that asymptotically the probability of short waiting times converges to 1 when the threshold time grows to infinity. This explains the reduction of ratio variance with increasing values of t_{thres}.

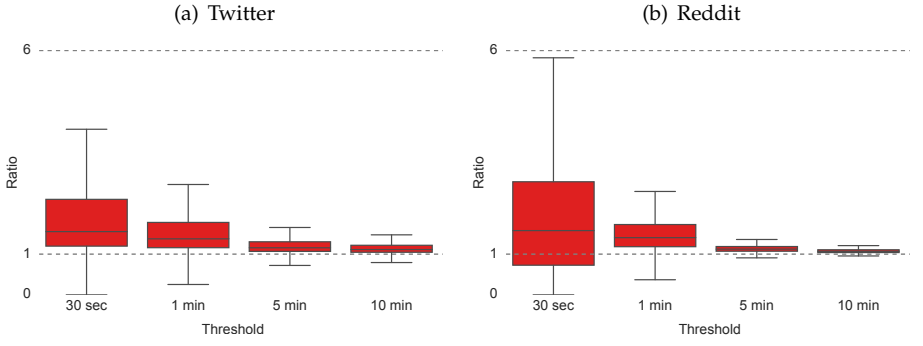


Fig. 6.4 Illustration of the dependence between two successive inter-event times. The box spreads around the median value q_2 (horizontal line) from the lower quartile (q_1) to the upper quartile (q_3) of the sample. The upper whisker extends below the value $q_3 + 1.5(q_3 - q_1)$ and the lower whisker extends above the value $q_1 - 1.5(q_3 - q_1)$.

6.4 A Markovian approach to incorporate memory

Users only generate events when they are connected on the social website and interested in communicating or reacting about a specific topic. One may think that knowing the last waiting time of a specific user could give us a hint about whether they will be again corresponding or tweeting in the near future. This activity property will be recorded by X_i. When a user generates events separated by a short waiting time ($X_i = \mathrm{S}$), we say that the user was in an *intensive* state during the inter-event time. Analogously, the user is considered as an *occasional* commenter during long waiting time ($X_i = \mathrm{L}$). An example of successive occasional and intensive states is given in Figure 6.5 for $t_{thres} = 1$ min.

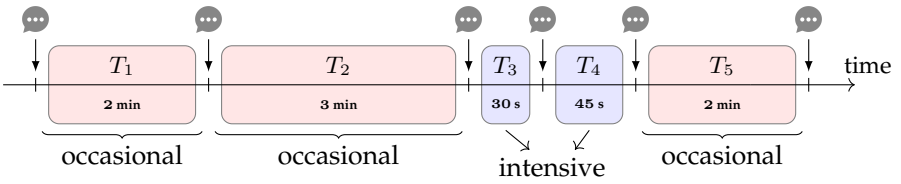


Fig. 6.5 Occasional and intensive states for a user that generates 6 arbitrar subsequent events. The threshold is set to 1 min in this particular example. Short waiting times are displayed in blue and long waiting times are displayed in red.

The choice of this structure has a twofold reason. First, using this one bit property X_i allows us to introduce a simple dependence structure, as will be described just below. Second, we plan to use a power-law distribution with density proportional to $\tau^{-\gamma}$ for some $\gamma > 1$. We remind that this can be used only on an interval bounded away from 0 if we want to make it well defined. Using this distribution only for *long* waiting times makes the model satisfy this constraint.

The key point of the model is the dependence structure we propose. We assume that the distribution at index $i + 1$ depends on the type X_i of the previous waiting time, but conditionally independent from anything else before. This 1-step memory system is displayed at Figure 6.6.

We have therefore a transition probability matrix for the type X_i :

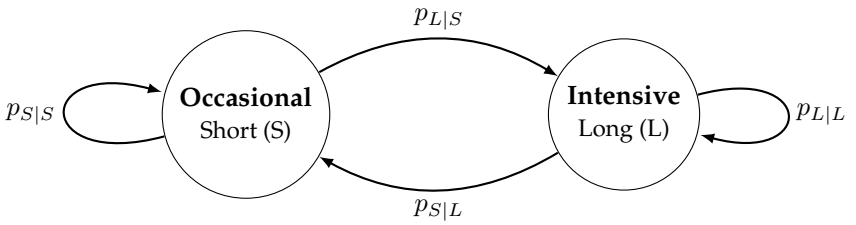


Fig. 6.6 Illustration of the underlying Markov-chain behind the waiting time distribution.

$$P = \begin{pmatrix} p_{S|S} & p_{S|L} \\ p_{L|S} & p_{L|L} \end{pmatrix}.$$

Here $p_{S|L} = P(X_{i+1} = \text{S}|X_i = \text{L})$ stands for the probability of observing a *short* waiting time after a *long* one for a specific user. Similar reasoning applies to the other entries of matrix P, from which we can calculate $p_S = P(X_i = \text{S})$ and $p_L = P(X_i = \text{L})$, the overall probabilities of having *short* or *long* waiting times. Observe that given $p_{S|S}$ and p_S for a specific user, we can compute all other probabilities concerning the process X_i.

We now incorporate probability density functions to model the variables T_i that are defined in a continuous state space. By focusing first on the long inter-event times, we denote by *stand-by* waiting times the variables $T_{i+1} > t_{thres}$ that follow an occasional state $X_i = \text{L}$. Similarly, the long waiting times that appear after intensive states $X_i = \text{S}$ are denoted *transition* waiting times. We introduce respectively the functions $\rho_{|L}$ and $\rho_{|S}$ as the density functions for the *stand-by* and *transition* waiting times. The motivation for considering two different densities is detailed at Section 6.5.4. We thus assume that the density might be different after a *short* and after a *long* waiting time. For these densities, we consider power-law distributions, giving them the form

$$\rho_{|S}(t) = \begin{cases} \frac{\gamma_S - 1}{t_{thres}} \left(\frac{t}{t_{thres}}\right)^{-\gamma_S} & \text{for } t \geq t_{thres} \\ 0 & \text{for } 0 \leq t < t_{thres} \end{cases}$$

$$\rho_{|L}(t) = \begin{cases} \frac{\gamma_L - 1}{t_{thres}} \left(\frac{t}{t_{thres}}\right)^{-\gamma_L} & \text{for } t \geq t_{thres} \\ 0 & \text{for } 0 \leq t < t_{thres} \end{cases}$$

For the short waiting times, we will simply consider a uniform probability density distribution f_U on $[0, t_{thres}]$. The continuous-state Markovian process

(denoted by MK) is then defined by the following conditional probability density function:

$$f_{T_{k+1}}(t_{k+1}|T_k = t_k) = \begin{cases} p_{S|L}\, f_U(t_{k+1}) + p_{L|L}\, \rho_{|L}(t_{k+1}) & \text{if } t_k \geq t_{thres} \\ p_{S|S}\, f_U(t_{k+1}) + p_{L|S}\, \rho_{|S}(t_{k+1}) & \text{if } 0 \leq t_k < t_{thres}\,. \end{cases} \tag{6.2-MK}$$

The model assumes that the threshold value t_{thres} is inherent for each social media. In other words, all the users interacting on a specific platform are assigned to one common threshold. A process with n users has therefore $4n+1$ degrees of freedom: one set of parameters $\left(p_S, p_{S|S}, \gamma_S, \gamma_L\right)$ per user and one threshold variable t_{thres}.

6.5 Assessing the time-dependent structure

Two improvements are analyzed in the following: the introduction of a threshold parameter that decomposes the waiting times into short and long types, and the consideration of a Markov time-dependent structure. We statistically evaluate the improvements by deriving two simplified models that operate as baselines. The two baseline models each incorporates one particular improvement.

6.5.1 Baseline models

Simpler models that do not incorporate time-dependence can be easily derived from the proposed Markovian process. First, we can drop the memory dependence by imposing $p_{S|S} = p_S$ as well as $\gamma_S = \gamma_L = \gamma$, creating a first baseline model denoted as the *Independent Threshold* model or IT. We now have one unique density function $\rho_{|L} = \rho_{|S} = \rho$ associated with the long waiting times. The process is specified by the following density function:

$$f_{T_{k+1}}(t_{k+1}) = p_S\, f_U(t_{k+1}) + p_L\, \rho\,(t_{k+1}) \tag{6.3-IT}$$

and is characterized by $2n+1$ degrees of freedom.

Furthermore, the inter-event times can also be fitted to one unique power-law density function without considering any threshold. In this case, we do not distinguish the short and the long waiting times. This can be easily achieved by the condition $p_S = 0$. The corresponding *Independent Power-law* model or IP is given by

$$f_{T_{k+1}}(t_{k+1}) = \rho\,(t_{k+1}) \tag{6.4-IP}$$

and associates different power-law exponents γ to each user, giving n degrees of freedom. Note that we use the power-law defined on 1 to ∞ . It is important to set a non-zero lower bound to have a proper probability distribution. The choice is appropriate since the recorded waiting times are bounded by the 1-second precision constraint of the timestamps.

The three models – IP, IT and MK– are nested models, in the sense that IP is a subset of IT, which in turn is a subset of MK.

6.5.2 Global assessment

The likelihood-ratio test statistics is appropriate to compare two models that are particular cases of one another [Lewis, Butler, and Gilbert (2011)]. The test takes into account the difference in complexity of the models and penalizes the more complex one.

We first compare the statistical significance whether the IT model should be preferred instead of the IP model on the whole population. Parameters are computed for each model by maximizing the likelihood over the n users. The LRTS is then performed by computing :

$$D_{cut} = 2\,(\log L_{IT} - \log L_{IP}).$$

We assess the distribution of this difference under the null hypothesis which specifies that the simpler model (IP) is the true model. The difference D_{cut} is always positive since the IP model is a particular case of the IT model. The variable D_{cut} is supposed to follow a χ-squared distribution with freedom equal to the number of extra parameters in the more refined model [Lewis, Butler, and Gilbert (2011)]. In this case, this makes $n + 1$ degrees of freedom. Similarly, we assess the significance of introducing a 1-step memory dependence by defining a second variable D_{mem}. This variable compares the Markovian model (MK) to the threshold model without any time dependence (IT):

$$D_{mem} = 2\,(\log L_{MK} - \log L_{IT}).$$

Following the same reasoning, the variable D_{mem} should follow a χ-squared distribution with $2n$ degrees of freedom under the hypothesis that the IT model is the true one.

We further define the critical value D_α^{df} related to a χ-squared distribution with degree df. Differences greater than this threshold ($D > D_\alpha^{df}$) reject the

	Cutting Short-Long			1-step Memory			
	D_{cut}	$D_{0.05}^{n+1}$	p-value	D_{mem}	$D_{0.05}^{2n}$	p-value	# pairs
Twitter	4 022 903	4 959	$< .001$	244 643	9 822	$< .001$	$N = 4\,802\,287$
Reddit	5 743 821	3 201	$< .001$	106 918	6 325	$< .001$	$N = 4\,172\,504$

Table 6.1 Significant improvements by introducing a threshold and by considering a 1-step memory process. The number of pairs specifies the amount of consecutive waiting times on which the likelihood is computed for both databases.

null hypothesis (i.e., that the simpler model is true) with an α-significance level. Table 6.1 displays the differences compared to critical values with level $\alpha = 0.05$. Clearly, we observe for both datasets that $D_{mem} \gg D_{0.05}^{2n}$, associated to significant p-values. This confirms our intuition that the Markovian model should be preferred to the Independent Threshold model. In turn, we have $D_{cut} \gg D_{0.05}^{n+1}$ which suggests that the IT model should be preferred to the classical power-law distribution. We remind that these conclusions take into account the difference of complexity of the compared model.

6.5.3 Performance per user

The analysis can be taken further by performing similar statistical tests at the user level. We now consider individual users and decide separately for each of them if the increase of model complexity is significantly improving the modeling of the observed waiting times.

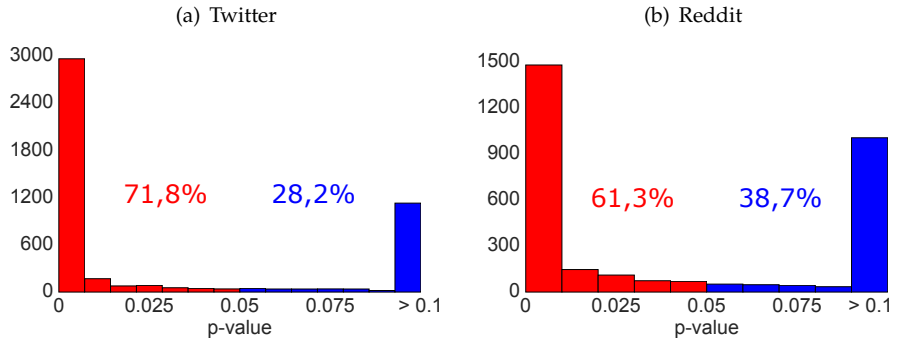


Fig. 6.7 Distribution of user p-values for the MK-IT comparison. Significant p-values favor the Markovian model compared to the Independent Threshold model.

We observe that the independent threshold model is significantly better at a level $\alpha = 0.05$ for every user than the classical approach. However, a similar conclusion does not hold for every user when the Markovian model is compared to the independent threshold approach. Figure 6.7 shows the distribution of the users' p-values that compares the Markovian model to the IT model, where low p-values favor the reject of the simpler model (i.e., the IT model). We observe that the p-values are significant ($\alpha = 0.05$) for a majority of users ($60\% - 70\%$) and that the Markovian model should be preferred over the IT model, even at the user level.

6.5.4 The power-law exponents

The Markovian model handles differently the long waiting times that directly follow short ones (the transition waiting times) from those following long ones (the stand-by waiting times). Both distributions are characterized by a power-law density with one identical lower bound t_{thres} but with different factors γ_S (transition exponent) and γ_L (stand-by exponent).

A strong motivation that brings us to use distinct power-law parameters comes from the link between activity[4] and power-law exponent. When γ increases, the expected waiting time decreases. In turn, the expected number of events in a fixed time frame increases. There is therefore a monotonous relation between the frequency of user events and the associated power-law factors. Choosing two different power-law factors translates dependence between the frequency of events and the current user state X_i.

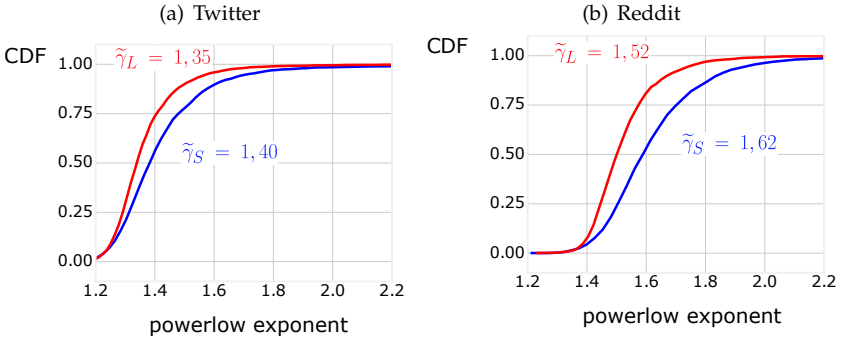


Fig. 6.8 Empirical distribution functions of power-law factors γ_S and γ_L. The Kolmogorov-Smirnov test rejects the equality of distributions with a p-value < 0.01. The graph displays the median values of both distributions.

[4]The frequency at which users generate events

We show at Figure 6.8 that the distribution of γ_S among the population is indeed different from the distribution of γ_L. A Kolmogorov-Smirnov test brings us to the conclusion that we can reject the equality of distribution for both datasets with a p-value < 0.01. The transition exponents, associated to an intensive state ($X_i = \text{S}$), appear clearly higher than the stand-by exponents. This is consistent with our intuition with the statement that higher exponents are associated with higher activity.

6.6 Concluding remarks

The goal of this last chapter was to propose an inspiring alternative for the standard model of human communication. The intuition of this modification is to highlight the dependence between subsequent waiting times. As a result, this allows stronger burstiness to be explained than what is captured by just declaring the waiting times to follow a power-law distribution. We investigate this new process with social media comments and we see convincing statistical evidence that this is the correct direction to go in order to improve the modeling of communication patterns. In particular, the new model fits significantly better even for a major proportion of singular users. Even more, for an aggregated comparison, the new model has a convincingly better fit.

This work gives importance to the simplicity of the model. We suggest that follow-up research activity investigates Markov chains with multiple states, or even Markov-2 processes where dependence extends to 2 steps in the past. Another possibility to incorporate interdependence would be to tailor self-exciting processes to the current situation. A primary candidate is the renowned Hawkes-process [Hawkes (1971)].
The aim of this last project was to build a mathematically sound model with a clean and simple dependence structure. The main idea was to direct attention to this dependence property that has to be exploited if we want a mature model describing human communication patterns.

Chapter 7

Conclusion

When we started the study of online opinions four years ago, we were initially attracted by the idea of investigating how personality traits can be linked to group influence. The whole project evolved rapidly towards the ambition to understand the dynamics of opinions in the social media. The main objective of our research became the modeling and predicting of judgment revision, and later the measurement of opinions on popular online networks.

A natural junction appears between Chapter 4 and Chapter 5. The first part is mainly based on *in-vitro* experiments. This gave us the opportunity to test numerous assumptions about human influence since the experimental conditions were largely controllable. The second part focuses on the observation of expressions and connexions in popular social networks. These *in-vivo* experiments allowed us to work under real-world conditions where participants are less tempted to alter their behavior, as it can be the case in controlled experiments.

Research outcomes

Our contributions are built around various themes: assessing unpredictable variations, modeling influence, extracting political positions and understanding communication patterns.

In the first part of the thesis, we introduced the concept of *intrinsic noise* through a variable that captures the part in human revision judgment that is composed of unpredictable variations. In particular, these variations limit the capability of conventional models to predict the evolution of human attitudes. We suggested a way to measure the intrinsic variation by replicating identical experimental conditions. To quantify opinion dynamics that are subject to social influence, we

carried out online experiments in which the participants had to estimate some quantities while receiving information about the other participants' opinions. In a first round, a participant expressed their isolated opinion x_i corresponding to their estimation in regard to the task. In the two subsequent rounds, the same participant was exposed to a set of opinions x_j and was then allowed to update their own opinion. Two types of games were designed : the *gauging game*, in which the participants evaluated color proportions, and the *counting game*, in which the participants had to guess the number of items displayed in a picture. The people taking part in this online crowdsourced study were involved in 3-round judgment revision games.
In the end, we discovered that about two thirds of the errors made by time-varying consensus models are due to these unpredictable variations. Our model is based on the *influenceability* $\alpha_i(t)$ of the participants, a factor representing the extent to which a participant incorporates external judgments. We assessed the predictions when the amount of training data is reduced, which corresponds to realistic settings where prior knowledge on individuals is scarce. Consensus models allow to draw interesting conclusions concerning the behavior of participants. In particular, we observed that a higher *influenceability* of agents promotes individual performances but undermines the collective performance. Yet, we found no evidence that the personality traits of participants are directly linked to their individual *influenceability*.

In the second part of the thesis, we focused on the measurement of people's opinion based on digital traces and we analyzed the communication patterns arising on social media. Capturing individual opinions on particular issues is not a straightforward process. Scholars assume the existence of an underlying concept denoted *ideology*. In the context of political elections, our results showed that valid ideological positions can be deduced for elites based solely on *network* and *textual* data publicly available on social media platforms. In particular, it is highly beneficial to group adjacent words by pairs (bigrams) when scaling from short and informal content. To test the validity of our estimates on ordinary citizens, we compared the estimates of these two methods to conventional *survey* data estimates. Network estimates are strongly correlated with survey ideological estimates, whereas the quality of textual estimates is more subtle. Nonetheless, the two types of estimates have been shown to be complementary, and combining them substantially improved the classification of citizens regarding their voting intentions.
We finally investigated online communication patterns by looking into the distribution of inter-event messages on Twitter and Reddit. We confirmed that subsequent waiting times are far from independent and that, consequently, independent power-law distributions are not sufficient to describe the observed

temporal structure. Based on the intuition that online users are either active or temporary inactive, we suggested two innovations that (a) separate waiting times into two categories (short and long waits), and (b) exhibit a Markovian dependent structure. Both innovations provide a significative improvement in the modeling of inter-arrival times.

Limitations and Future Work

Follow-up researchers should be aware of the following limitations and challenges inherent to the present work.

Multidimensionality Our work considers unidimensional opinions and ideologies. However, they are many situations in which the unidimensional assumption does not hold. We already saw that ideological positions in the context of Quebec 2014 can hardly be described by one unique dimension. As another example, Lorenz (2007) gives a framework in which the opinions of agents are compared to a budget plan proposal. He models multidimensional opinions by considering agents that have to allocate a fixed amount of money to a number $d > 1$ of projects. The models introduced by Deffuant, Neau, Amblard et al. (2000) as well as the model suggested by Hegselmann and Krause (2002) are two examples that handle multidimensional opinions. Both approaches consider an opinion space $S \in \mathbb{R}^d$, where S is usually compact and convex.

The distance between opinions is a predominant feature in the study of opinions. Most models assume that agents tend to update their opinion in function of its distance from the opinion of others. For instance, the network scaling approach suggested by Barberá (2015) models the connexions between agents as a function of the ideological distances. Opinion distances in $\mathbb{R}^d$ are generally induced by a norm. While there exists only a single distance (induced by a norm) in the unidimensional case – i.e., the Euclidean distance – they are multiple possible norms when we expand the analysis to multidimensional opinions. Beyond that, multidimensionality also generates visualization challenges. As an example, we remind that agents are visually exposed to the others' opinion during the crowdsourcing experiments. While a bidimensional visualization seems tractable, it becomes difficult for $d \geq 3$.

Unidimensional ideologies are a standard choice in the political science literature. However, most of the scaling models can be extended to a multidimensional analysis. For instance, extracting bidimensional ideologies from network information only increases the number of parameters from $2(n + m)$ to $3(n + m)$. In the textual approach, Slapin and Proksch (2008) suggest classifying the documents (or in our case, the tweets) into predefined categories. For

instance, an economic dimension can be extracted by filtering all the sentences containing economic references.

Discrete opinions We only include continuous opinion dynamic models in the present work. However, opinions are often expressed in terms of binary choices or Likert scales. Continuous state experiments are easily converted into discrete opinion studies (e.g., *does the picture display more than* 200 *items?*). Discrete choice models include the rumor [Dodds and Watts (2004)] and threshold models [Granovetter (1978); Watts (2002)]. In the latter, an individual modifies their actions when a certain proportion of their neighbors does so.
A recent model bridges these two bulks of work by considering the social influence of discrete actions on continuous opinions which itself results in an individual action [Martins (2008)]. This is known as the CODA model. It appears that this model also naturally leads to cascade of behaviors over a social network [Chowdhury, Morarescu, Martin et al. (2016)]. It would be interesting to see how the threshold parameter in Granovetter's model may be expressed in terms of the initial opinion of individuals in the CODA model : individuals with an opinion close to the boundary leading to either of the two discrete actions would have a lower threshold for action change.

Continuous time and additional rounds The judgment revision models are expressed in discrete time since the related experiments only allow the players to update their opinions at specific moments (i.e., rounds) and a limited number of times (in this work, two updates), regulated by a timer. We targeted two research prospects: analyzing the evolution of opinions after longer periods of time by including additional rounds, and allowing a real-time update.
We relinquished the idea of integrating extra rounds with regard to the already lengthy duration of the experiment ($\simeq$ 45 min). In addition, such a procedure require to incorporate an additional parameter per extra round ($\alpha_i(t_k)$) for each participant, which increases the risk of overfitting. Alternatively, the increase of complexity is partially alleviated when the time variation is embedded in a decaying parameter, as proposed in Section 4.6 of Chapter 4. However, the regression process does not always lead to one unique solution.
For technical reasons, we did not launch an experiment with a real-time opinion update but we highly recommend future researchers investigate these directions.

Controlling opinions towards a desired output One of the main contributions is the development of a framework that assesses the predictive power of opinion dynamic models. Beyond the prediction perspective, one natural follow-up research would be to question the ability of these models to control

the group opinion towards a desired output. To the future researchers having an interest in control issues, we would recommend to study the problem in a twofold way, that is:

- either by designing experiments where the researcher controls the existence or the nature of the connexions between agents. This goes back to the idea of modeling influence with linear switched systems $x(t+1) = A_{\sigma(t)}\, x(t)$ (see for instance [Daafouz, Riedinger, and Iung (2002)] or more recently [Philippe, Essick, Dullerud et al. (2016)]);
- or by directly hiding virtual players in the experiment with controllable opinions $u(t)$. In this case, the problem aims at designing controllers for systems of the type $x(t+1) = A(t)\, x(t) + B(t)\, u(t)$.

Various controlling tasks may be used depending on the research purposes. Classically, we may want to control opinions towards a fixed value x^*. Another option would be to test the group cohesion by trying to split the players into multiple clusters.

Micro-level attributes In the crowdsourcing experiment, we investigated how microlevel attributes affect influence within a collective. We observed that the link between personality traits and human influence was not conclusive. However, micro-level elements such as personality, trust, motivation, demographics and intelligence are expected to heavily influence group dynamics [Kozlowski and Bell (2003)] and may thus be some of the enabling factors of collective intelligence [Salminen (2012)]. Thus, future research should clearly focus on the relationships between these individual attributes and the group outcome. This would for instance answer the need for business to know how to establish teams [Lawler, Mohrman, and Ledford (1998)]. As all these features will impact the decision outcome in some ways, we encourage researchers to investigate the following questions: who should compose the collective ? What should be the interaction pattern? Should certain individuals have more importance than others or should the decision process be completely flat ? Scholars already have cues on which attributes affect the outcome in collective decision-making. For instance, it has been shown that attributes such as the general mental ability [Bell (2007); Devine and Philips (2001)] or (and less obviously) social sensitivity [Woolley, Chabris, Pentland et al. (2010)] enhance the performance of groups. When polling individuals, the average of opinions is generally more accurate in groups of higher personality diversity [Jain, Bearden, and Filipowicz (2011)]. Moreover, it has been shown that collective intelligence works even with non-experts [Hong and Page (2004); Lakhani, Jeppesen, Lohse et al. (2007)].

Social media for election forecast Online information is a useful resource to analyze the political behavior of citizens. Despite the prevailing skepticism about social media data, we have seen that machine learning approaches are likely to generate accurate predictions on the voting intentions of politically involved users. Nevertheless, surveys still remain the reference tool for election forecasts on a global scale. Indeed, social media data are far less representative of the potential voters than survey data sets [Diaz, Gamon, Hofman et al. (2016)].
Scholars started to develop models to predict socio-economic variables such as gender, age, and income, in order to recover the representativeness of the samples. In fact, measuring these variables seems far easier than estimating someone's ideology. Therefore, recovering the representativeness of these highly non-representative samples is a promising perspective, and has already been successfully applied to evaluate public opinion on a broad range of topics (Wang, Rothschild, Goel et al., 2015).
In addition to the issue of online bias, there is uncertainty about the future of these social platforms. While the scaling methods presented here are expected to deliver accurate estimates for Twitter, there is no guarantee that the platform will remain active over the years.

Politically inactives While the methods are successful at scaling individual-level ideologies and predicting vote intentions, this is done while paying a cost in high filtering criteria. These criteria directly affect the size of the user samples from which we can estimate political ideologies. For instance, our method does not provide ideological estimations for citizens who do not follow any political elite. In addition, only citizens that posted messages during election periods were retained for the analysis. Reducing the filter thresholds (in order to inflate the sample size) would inevitably decrease the quality of the estimates: the reduction of the β parameter will generate additional N-grams mainly composed of noisy words that do not capture any political content.
Future research should thus explore and compare these estimates during time periods other than election campaigns. In a more restrictive case, one could wonder whether valid estimates can possibly be computed for politically inactive users. To tackle this question, a path forward would be to apply a user-based collaborative filter. Citizens who would not pass the required filtering criteria would be compared to other politically active users. Their ideology would then be inferred from their neighbors' ideologies. Therefore, a measure of similarity would have to be defined between active and inactive citizens.

In-vivo opinion dynamics We have seen how opinion dynamic models can derive estimates of future opinions from the social agents' past behaviors.

There are still a lot of issues to address before extending our in-vitro results to real social networks.
The first limitations come from the fact that opinions were generated from simple tasks (games). Even if the results are replicable for two slightly different kinds of games, we cannot be sure that our models will still be predictive in the context of daily opinions (such as political attitudes or consumers' opinions). In addition, we only considered a simplified form of interaction in this work. Indeed, in the in-vitro experiments, agents are completely connected in small groups (the underlying graph is the complete graph K_6) and are directly and simultaneously exposed to the numerical opinions of the group. On real networks, members are rather influenced by being exposed locally to public posts or private messages that appear to them randomly over time (as detailed in the last chapter). Finally (and fortunately), interactions still happen outside of the online world. These face-to-face interactions are not recorded by social media, and they undoubtedly increase the level of unpredictability when predicting from online data.

Closing remarks

Every day, social media platforms generate massive amounts of data. Messages and network data are available on a large scale, not only for long periods of time on a daily basis, but also for different strata of the population. These data can offer fine-grained estimates that can increase the scope of opinion research to study changes of opinions over short periods of time, and can also foster the study of these effects on specific subgroups of agents or minorities.

Analogously, we believe that crowdsourcing will be increasingly used in future psychosocial experiments. Computer-mediated experiments as a whole are indeed more tunable than controlled laboratory environments. Virtual reality may allow us to purposely mimic some aspects of face-to-face interaction. In addition, communication can be more easily structured and the interaction network can be readily modified. Interactions may stretch over different periods of time, being non necessarily simultaneous.

In either case, these new data sources offer promising opportunities to put to the test numerous theories on human behavior that previously remained theoretical.

Bibliography

Abbott, D (2013). The reasonable ineffectiveness of mathematics [point of view]. *Proceedings of the IEEE*, 101(10):2147–2153.

Adcock, R (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Association*, 95(03):529–546.

Aledavood, T, López, E, Roberts, SG, Reed-Tsochas, F, Moro, E, et al. (2015). Daily rhythms in mobile telephone communication. *PLOS ONE*, 10(9):e0138098.

Ansolabehere, S, Rodden, J, and Snyder, JM (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, 102(02):215–232.

Aral, S, Muchnik, L, and Sundararajan, A (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549.

Aral, S and Walker, D (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Ariely, D, Tung Au, W, Bender, RH, Budescu, DV, Dietz, CB, et al. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, 6(2):130.

Azen, R and Budescu, DV (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2):129.

Barabási, AL (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211.

Barber, D (2012). Factor analysis. In Barber, D, editor, *Bayesian reasoning and machine learning*, pages 462–478. Cambridge University Press.

Barberá, P (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Bavelas, A (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Beauchamp, N (2016). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*.

Becker, GS (2013). *The economic approach to human behavior*. University of Chicago press.

Bell, ST (2007). Deep-level composition variables as predictors of team performance: a meta-analysis. *Journal of applied psychology*, 92(3):595.

Berger, RL (1981). A necessary and sufficient condition for reaching a consensus using degroot's method. *Journal of the American Statistical Association*, 76(374):415–418.

Bergman, MM (1998). A theoretical note on the differences between attitudes, opinions, and values. *Swiss Political Science Review*, 4(2):81–93.

Bernardes, AT, Stauffer, D, and Kertész, J (2002). Election results and the sznajd model on barabasi network. *The European Physical Journal B-Condensed Matter and Complex Systems*, 25(1):123–127.

Bessi, A, Coletto, M, Davidescu, GA, Scala, A, Caldarelli, G, et al. (2015). Science vs conspiracy: collective narratives in the age of misinformation. *PloS one*, 10(2):02.

Bishop, CM et al. (2006). *Pattern recognition and machine learning*, volume 1. Springer New York.

Blondel, V, Hendrickx, J, and Tsitsiklis, J (2009). On Krause's multi-agent consensus model with state-dependent connectivity. *IEEE Transactions on Automatic Control*, 54(11):2586–2597.

Blondel, VD, Decuyper, A, and Krings, G (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10.

Blumenstock, JE, Gillick, D, and Eagle, N (2010). Who's calling? demographics of mobile phone use in rwanda. *Transportation*, 32:2–5.

Bonaccio, S and Dalal, RS (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151.

Bond, R and Messing, S (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review*, 109(01):62–78.

Bond, RM, Fariss, CJ, Jones, JJ, Kramer, AD, Marlow, C, et al. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

Bonica, A (2013). Ideology and interests in the political marketplace. *American Journal of Political Science*, 57(2):294–311.

Bonica, A (2014). Mapping the ideological marketplace. *American Journal of Political Science*, 58(2):367–386.

Brown, PF, Desouza, PV, Mercer, RL, Pietra, VJD, and Lai, JC (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Budescu, DV and Rantilla, AK (2000). Confidence in aggregation of expert opinions. *Acta psychologica*, 104(3):371–398.

Campbell, A, Converse, PE, Miller, W, and Stokes, D (1960). *The American Voter*. Wiley, New-York.

Carpini, MXD and Keeter, S (1997). *What Americans know about politics and why it matters*. Yale University Press.

Carroll, R, Lewis, JB, Lo, J, Poole, KT, and Rosenthal, H (2009). Measuring bias and uncertainty in dw-nominate ideal point estimates via the parametric bootstrap. *Political Analysis*, 17(3):261–275.

Caruso, F and Castorina, P (2005). Opinion dynamics and decision of vote in bipolar political systems. *International Journal of Modern Physics C*, 16(09):1473–1487.

Caton, S, Hall, M, and Weinhardt, C (2015). How do politicians use facebook? an applied social observatory. *Big Data & Society*, 2(2):2053951715612822.

Centola, D (2010). The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197.

Ceron, A (2017). Intra-party politics in 140 characters. *Party Politics*, 23(1):7–17.

Chacoma, A and Zanette, DH (2015). Opinion formation by social influence: From experiments to modeling. *PloS one*, 10(10):e0140406.

Chatterjee, S and Seneta, E (1977). Towards consensus: Some convergence theorems on repeated averaging. *Journal of Applied Probability*, 14(01):89–97.

Chowdhury, N, Morarescu, IC, Martin, S, and Srikant, S (2016). Continuous opinions and discrete actions in social networks: a multi-agent system approach. *arXiv preprint arXiv:1602.02098*.

Christakis, NA and Fowler, JH (2008). The collective dynamics of smoking in a large social network. *New England Journal of Medicine*, 358(21):2249–2258. doi: 10.1056/NEJMsa0706154. PMID: 18499567.

Clauset, A, Shalizi, CR, and Newman, ME (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Clearwater, SH, Huberman, BA, and Hogg, T (1991). Cooperative solution of constraint satisfaction problems. *Science*, 254(5035):1181.

Clinton, J, Jackman, S, and Rivers, D (2004). The statistical analysis of roll call data. *American Political Science Review*, 98(02):355–370.

Colic-Peisker, V (2016). Ideology and utopia: Historic crisis of economic rationality and the role of public sociology. *Journal of Sociology*, page 1440783316630114.

Conover, PJ and Feldman, S (1981). The origins and meaning of liberal/conservative self-identifications. *American Journal of Political Science*, pages 617–645.

Converse, PE (1964). The nature of belief systems in mass publics. In Apter, D, editor, *Ideology and Discontent*. Free Press of Glencoe, New-York.

Cross, R, Borgatti, SP, and Parker, A (2002). Making invisible work visible: Using social network analysis to support strategic collaboration. *California management review*, 44(2):25–46.

Daafouz, J, Riedinger, P, and Iung, C (2002). Stability analysis and control synthesis for switched systems: a switched lyapunov function approach. *IEEE transactions on automatic control*, 47(11):1883–1887.

Davis, OA, Hinich, MJ, and Ordeshook, PC (1970). An expository development of a mathematical model of the electoral process. *American Political Science Review*, 64(02):426–448.

Davis-Stober, CP, Budescu, DV, Dana, J, and Broomell, SB (2014). When is a crowd wise? *Decision*, 1(2):79.

Decuyper, A, Rutherford, A, Wadhwa, A, Bauer, JM, Krings, G, et al. (2014). Estimating food consumption and poverty indices with mobile phone data. *arXiv preprint arXiv:1412.2595*.

Deffuant, G, Neau, D, Amblard, F, and Weisbuch, G (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(1–4):87–98. doi: 10.1142/S0219525900000078.

DeGroot, M (1974). Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121.

Dellarocas, C (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, 49(10):1407–1424.

Devine, DJ and Philips, JL (2001). Do smarter teams do better a meta-analysis of cognitive ability and team performance. *Small Group Research*, 32(5):507–532.

Diaz, F, Gamon, M, Hofman, JM, Kıcıman, E, and Rothschild, D (2016). Online and social media data as an imperfect continuous panel survey. *PloS one*, 11(1):e0145406.

Dodds, PS and Watts, DJ (2004). Universal behavior in a generalized model of contagion. *Physical review letters*, 92(21):218701.

Downs, A (1957). *An economic theory of democracy*. New York: Addison Wesley.

Ellis, C and Stimson, JA (2012). *Ideology in America*. Cambridge: Cambridge University Press.

Feldman, S and Johnston, C (2014). Understanding the determinants of political ideology: Implications of structural complexity. *Political Psychology*, 35(3):337–358.

Feng, B and MacGeorge, EL (2006). Predicting receptiveness to advice: Characteristics of the problem, the advice-giver, and the recipient. *Southern Communication Journal*, 71(1):67–85.

Fischer, I and Harvey, N (1999). Combining forecasts: What information do judges need to outperform the simple average? *International journal of forecasting*, 15(3):227–246.

Forsyth, DR (2009). *Group dynamics*. Cengage Learning.

Fortunato, S and Castellano, C (2007). Scaling and universality in proportional elections. *Physical Review Letters*, 99(13):138701.

Fortunato, S, Latora, V, Pluchino, A, and Rapisarda, A (2005). Vector opinion dynamics in a bounded confidence consensus model. *International Journal of Modern Physics C*, 16(10):1535–1551.

French, J (1956). A formal theory of social power. *Psychological Review*, 63:181–194.

Friedkin, NE and Johnsen, EC (1990). Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206.

Galton, F (1907). Vox populi (the wisdom of crowds). *Nature*, 75:450–451.

Gerring, J (1997). Ideology: A definitional analysis. *Political Research Quarterly*, 50(4):957–994.

Gino, F (2008). Do we listen to advice just because we paid for it? the impact of advice cost on its use. *Organizational Behavior and Human Decision Processes*, 107(2):234–245.

Gino, F and Schweitzer, ME (2008). Blinded by anger or feeling the love: how emotions influence advice taking. *Journal of Applied Psychology*, 93(5):1165.

Godbout, JF and Høyland, B (2011). Legislative voting in the canadian parliament. *Canadian Journal of Political Science*, 44(02):367–388.

Goh, KI and Barabási, AL (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, 81(4):48002.

Gosling, SD, Rentfrow, PJ, and Swann, WB (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Granovetter, M (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83:1420–1443.

Granovetter, MS (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Grimmer, J and Stewart, BM (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*.

Gulliver, A (2014). National's plan will 'stabilise house prices', http://www.stuff.co.nz/national/politics/10421509/nationals-plan-will-stabilise-house-prices.

Harman, HH (1960). Modern factor analysis.

Harries, C, Yaniv, I, and Harvey, N (2004). Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, 17(5):333–348.

Harvey, N and Fischer, I (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70(2):117–133.

Hastie, R, Penrod, S, and Pennington, N (1983). *Inside the jury*. The Lawbook Exchange, Ltd.

Hawkes, AG (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, pages 83–90.

Haynes, AA and Pitts, B (2009). Making an impression: New media in the 2008 presidential nomination campaigns. *PS: Political Science & Politics*, 42(01):53–58.

Hegselmann, R and Krause, U (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Henrich, J, Heine, SJ, and Norenzayan, A (2010). The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Hoffman, MD and Gelman, A (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hong, L and Page, SE (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, 101(46):16385–16389.

Hong, L and Page, SE (2008). Some microfoundations of collective wisdom. *Collective Wisdom*, pages 56–71.

Horowitz, IA, ForsterLee, L, and Brolly, I (1996). Effects of trial complexity on decision making. *Journal of applied psychology*, 81(6):757.

Hsu, CW, Chang, CC, Lin, CJ, et al. (2003). A practical guide to support vector classification.

Huang, YM and Du, SX (2005). Weighted support vector machine for classification with uneven training class sizes. In *2005 International Conference on Machine Learning and Cybernetics*, volume 7, pages 4365–4369. IEEE.

Iglewicz, B and Hoaglin, DC (1993). *How to detect and handle outliers*, volume 16. ASQC Quality Press Milwaukee (Wisconsin).

Iñiguez, G, Török, J, Yasseri, T, Kaski, K, and Kertész, J (2014). Modeling social dynamics in a collaborative environment. *EPJ Data Science*, 3(1):1–20.

Jain, K, Bearden, JN, and Filipowicz, A (2011). Diverse personalities make for wiser crowds: How personality can affect the accuracy of aggregated judgments.

Jessee, S (2016). (How) can we estimate the ideology of citizens and political elites on the same scale? *American Journal of Political Science*.

Jessee, SA (2009). Spatial voting in the 2004 presidential election. *American Political Science Review*, 103(01):59–81.

Jessee, SA (2012). *Ideology and spatial voting in American elections*. Cambridge: Cambridge University Press.

Jewitt, CE and Goren, P (2016). Ideological structure and consistency in the age of polarization. *American Politics Research*, 44(1):81–105.

Johansen, A (2004). Probing human response times. *Physica A: Statistical Mechanics and its Applications*, 338(1):286–291.

Kam, CJ (2009). *Party discipline and parliamentary politics*. Cambridge University Press.

Karsai, M, Kivelä, M, Pan, RK, Kaski, K, Kertész, J, et al. (2011). Small but slow world: How network topology and burstiness slow down spreading. *Physical Review E*, 83(2):025102.

Kemeny, JGJG (1972). Mathematical models in the social sciences. Technical report.

Kersten, D and Yuille, A (2003). Bayesian models of object perception. *Current opinion in neurobiology*, 13(2):150–158.

Kittur, A, Chi, EH, and Suh, B (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

Kovanen, L, Kaski, K, Kertész, J, and Saramäki, J (2013). Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, 110(45):18070–18075.

Kozlowski, SW and Bell, BS (2003). Work groups and teams in organizations. *Handbook of psychology*.

Krackhardt, D (1996). Social networks and the liability of newness for managers. *Trends in organizational behavior*, 3:159–173.

Krause, U (2000). A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, pages 227–236.

Krause, U and Stöckler, M (1997). Modellierung und simulation von dynamiken mit vielen interagierenden akteuren. *Modus. Universität Bremen*.

Kritzer, HM (1978). Ideology and american political elites. *Public Opinion Quarterly*, 42(4):484–502.

Kruskal, JB and Wish, M (1978). *Multidimensional scaling*, volume 11. Sage.

Lakhani, KR, Jeppesen, LB, Lohse, PA, Panetta, JA, et al. (2007). *The value of openess in scientific problem solving*. Division of Research, Harvard Business School.

Lampos, V, Preotiuc-Pietro, D, and Cohn, T (2013). A user-centric model of voting intention from social media. In *ACL (1)*, pages 993–1003.

Lane, RE (1962). Political ideology: why the american common man believes what he does.

Larrick, RP and Soll, JB (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science*, 52(1):111–127.

Lawler, EE, Mohrman, SA, and Ledford, GE (1998). *Strategies for high performance organizations: The CEO report: Employee involvement, TQM, and reengineering programs in Fortune 1000 corporations*. Jossey-Bass Publishers San Francisco.

Lewis, F, Butler, A, and Gilbert, L (2011). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2):155–162.

Van der Linden, D, te Nijenhuis, J, and Bakker, AB (2010). The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327.

Lorenz, J (2007). Continuous opinion dynamics of multidimensional allocation problems under bounded confidence: More dimensions lead to better chances for consensus. *arXiv preprint arXiv:0708.2923*.

Lorenz, J (2008). Fostering consensus in multidimensional continuous opinion dynamics under bounded confidence. *Managing complexity: insights, concepts, applications*, pages 321–334.

Lorenz, J, Rauhut, H, Schweitzer, F, and Helbing, D (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Lukin, S and Walker, M (2013). Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40. Citeseer.

Luskin, RC (1990). Explaining political sophistication. *Political Behavior*, 12(4):331–361.

Ma, WJ, Beck, JM, Latham, PE, and Pouget, A (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438.

Malmgren, RD, Stouffer, DB, Motter, AE, and Amaral, LA (2008). A poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences*, 105(47):18153–18158.

Mannes, AE (2009). Are we wise about the wisdom of crowds? the use of group judgments in belief revision. *Management Science*, 55(8):1267–1279.

Martin, S and Girard, A (2013). Continuous-time consensus under persistent connectivity and slow divergence of reciprocal interaction weights. *SIAM Journal on Control and Optimization*, 51(3):2568–2584.

Martin, S, Saalfeld, T, Strøm, KW, Slapin, JB, and Proksch, SO (2014). Words as data: Content analysis in legislative studies.

Martins, AC (2008). Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(04):617–624.

Mason, W, Conrey, F, and Smith, E (2007). Situating social influence processes: dynamic, multidirectional flows of influence within social networks. *Pers Soc Psychol Rev*, 11(3):279–300.

Mason, W and Suri, S (2012). Conducting behavioral research on amazon?s mechanical turk. *Behavior research methods*, 44(1):1–23.

Mavrodiev, P, Tessone, CJ, and Schweitzer, F (2013). Quantifying the effects of social influence. *Scientific reports*, 3.

McGuire, WJ (1969). The nature of attitudes and attitude change. *The handbook of social psychology*, 3(2):136–314.

Ming-Sheng, S, Guan-Xiong, C, Shuang-Xing, D, Bing-Hong, W, and Tao, Z (2010). Interest-driven model for human dynamics. *Chinese Physics Letters*, 27(4):048701.

Monroe, BL and Maeda, K (2004). Talk's cheap: Text-based estimation of rhetorical ideal-points. In *annual meeting of the Society for Political Methodology*, pages 29–31.

Moussaïd, M, Kämmer, JE, Analytis, PP, and Neth, H (2013). Social influence and the collective dynamics of opinion formation. *PloS one*, 8(11):e78433.

Muchnik, L, Aral, S, and Taylor, SJ (2013). Social influence bias: A randomized experiment. *Science*, 341(6146):647–651.

Newman, ME (2002). Assortative mixing in networks. *Physical review letters*, 89(20):208701.

Niermann, S (2007). Testing for linearity in simple regression models. *AStA Advances in Statistical Analysis*, 91(2):129–139.

Olfati-Saber, R, Fax, JA, and Murray, RM (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.

Omnicore (2017). Twitter by the numbers: Stats, demographics & fun facts, https://www.omnicoreagency.com/twitter-statistics. *Salman Aslam*.

Oskamp, S and Schultz, PW (2005). *Attitudes and opinions*. Psychology Press.

Page, BI and Shapiro, RY (2010). *The rational public: Fifty years of trends in Americans' policy preferences*. University of Chicago Press.

Philippe, M, Essick, R, Dullerud, GE, and Jungers, RM (2016). Stability of discrete-time switching systems with constrained switching sequences. *Automatica*, 72:242–250.

Poole, KT and Rosenthal, H (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Poole, KT and Rosenthal, H (1987). Analysis of congressional coalition patterns: A unidimensional spatial model. *Legislative Studies Quarterly*, pages 55–75.

Poole, KT and Rosenthal, H (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Poole, KT and Rosenthal, H (2001). D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting. *Legislative Studies Quarterly*, pages 5–29.

Pope, DG (2009). Reacting to rankings: evidence from "america's best hospitals". *Journal of health economics*, 28(6):1154–1165.

Rabinowitz, G and Macdonald, SE (1989). A directional theory of issue voting. *American Political Science Review*, 83(01):93–121.

Rand, W and Rust, RT (2011). Agent-based modeling in marketing: Guidelines for rigor. *International Journal of Research in Marketing*, 28(3):181–193.

Read, J, Pfahringer, B, Holmes, G, and Frank, E (2011). Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359.

Rocha, LE, Decuyper, A, and Blondel, VD (2013). Epidemics on a stochastic model of temporal network. In *Dynamics On and Of Complex Networks, Volume 2*, pages 301–314. Springer.

Rummel, RJ (1967). Understanding factor analysis. *Journal of Conflict Resolution*, pages 444–480.

Salathé, M, Kazandjieva, M, Lee, JW, Levis, P, Feldman, MW, et al. (2010). A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025.

Salganik, MJ, Dodds, PS, and Watts, DJ (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

Salminen, J (2012). Collective intelligence in humans: A literature review. *arXiv preprint arXiv:1204.3401*.

Schrah, GE, Dalal, RS, and Sniezek, JA (2006). No decision-maker is an island: integrating expert advice with information acquisition. *Journal of Behavioral Decision Making*, 19(1):43–60.

See, KE, Morrison, EW, Rothman, NB, and Soll, JB (2011). The detrimental effects of power on confidence, advice taking, and accuracy. *Organizational Behavior and Human Decision Processes*, 116(2):272–285.

Shalizi, CR and Thomas, AC (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239.

Shaw, ME (1971). Group dynamics: The psychology of small group behavior.

Simmel, G (1908). Sociology: investigations on the forms of sociation. *Duncker & Humblot, Berlin Germany*.

Simon, HA (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 69(1):99–118.

Slapin, JB and Proksch, SO (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Smola, A and Vishwanathan, S (2008). Introduction to machine learning. *Cambridge University, UK*, 32:34.

Sniezek, JA, Schrah, GE, and Dalal, RS (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, 17(3):173–190.

Soll, JB and Larrick, RP (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3):780.

Steyvers, M, Griffiths, TL, and Dennis, S (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, 10(7):327–334.

Stimson, JA (2004). *Tides of consent: How public opinion shapes American politics*. Cambridge: Cambridge University Press.

Stokes, DE (1963). Spatial models of party competition. *American political science review*, 57(02):368–377.

Struppeck, T (2014). Combining estimates. *Casualty Actuarial Society E-Forum*, 2:1–14.

Takaguchi, T, Sato, N, Yano, K, and Masuda, N (2012). Importance of individual events in temporal networks. *New Journal of Physics*, 14(9):093003.

Temporão, M, Vande Kerckhove, C, Hendrickx, JM, and Dufresne, Y (2016). On the validity of social media data for public opinion research. *Submitted to Political Analysis*.

Toff, BJ and Kim, YM (2013). Words that matter: Twitter and partisan polarization.

Török, J, Iñiguez, G, Yasseri, T, San Miguel, M, Kaski, K, et al. (2013). Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters*, 110(8):088701.

Valente, TW (2012). Network interventions. *Science*, 337(6090):49–53.

Vande Kerckhove, C, Martin, S, Gend, P, Rentfrow, PJ, Hendrickx, JM, et al. (2016). Modelling influence and opinion evolution in online collective behaviour. *PloS one*, 11(6):e0157685.

Vazquez, A (2007). Impact of memory on human dynamics. *Physica A: Statistical Mechanics and its Applications*, 373:747–752.

Vul, E and Pashler, H (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, 19(7):645–647.

Walker, JD (2016). *Wisdom of the crowd mechanisms*. Ph.D. thesis, University of Pittsburgh.

Wang, LX and Mendel, JM (2016). Fuzzy opinion networks: A mathematical framework for the evolution of opinions and their uncertainties across social networks. *IEEE Transactions on Fuzzy Systems*, 24(4):880–905.

Wang, W, Rothschild, D, Goel, S, and Gelman, A (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Watts, DJ (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771.

White, LA (1943). Sociology, physics and mathematics. *American Sociological Review*, pages 373–379.

Whitehorn, A (2015). Green party of canada. http://www.thecanadianencyclopedia.ca/en/article/green-party-of-canada. *Digital Marketing*.

Woolley, AW, Chabris, CF, Pentland, A, Hashmi, N, and Malone, TW (2010). Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688.

Yaniv, I (2004a). The benefit of additional opinions. *Current directions in psychological science*, 13(2):75–78.

Yaniv, I (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1):1–13.

Yaniv, I and Kleinberger, E (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, 83(2):260–281.

Yaniv, I and Milyavsky, M (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, 103(1):104–120.

Zephoria (2017). The top 20 valuable facebook statistics (updated may 2017). https://zephoria.com/top-15-valuable-facebook-statistics. *Digital Marketing*.

Zhou, T, Kiet, HAT, Kim, BJ, Wang, BH, and Holme, P (2008). Role of activity in human dynamics. *EPL (Europhysics Letters)*, 82(2):28002.

Appendix A

The crowdsourcing experiment

The present section describes the experiment interface. A freely accessible single-player version of the games was also developed to provide a first-hand experience of the games. In the single-player version, participants are exposed to judgments stored on our database obtained from real participants in previous games. The interface and the timing of the single-player version is the same as the version used in the control experiment. The only difference is that the freely accessible version does not involve redundant games and provides accuracy feedback to the participants.

A.1 On the crowdsourcing platform

In the multiplayer version which was used for the uncontrolled experiment, the participants came from the *CrowdFlower*® external crowdsourcing platform where they received the URL of the experiment login page along with a key code to be able to login. The ad we posted on CrowdFlower was as follows :

> ***Estimation game regarding color features in images***
> *You will be making estimations about features in images. Beware that this game is a 6-player game. If not enough people access the game, you will not be able to start and get rewarded. To start the game : click on <estimation-game> and login using the following information :*
> *login : XXXXXXXX*
> *password: XXXXXXXX*
> *You will receive detailed instruction there. At the end of the game you will*

receive a reward code which you must enter below in order to get rewarded.

The participants were told they will be given another key code at the end of the experiment which they had to use to get rewarded on the crowdsourcing platform, this forced the participants to finish the experiment if they wanted to obtain a payment.

A.2 On the personal website

When the participants arrived on the experiment login page, they chose a login name and password so they could come back using the same login name if they wanted to for another experiment (see Figure A.1).

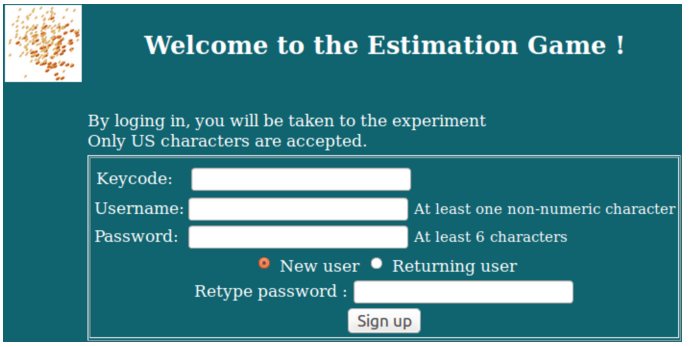
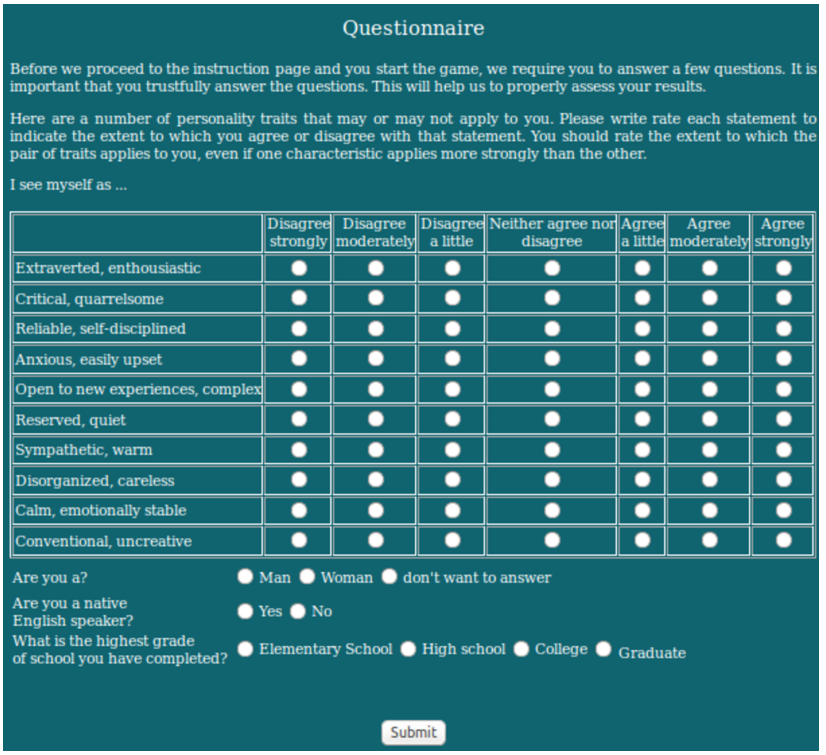


Fig. A.1 Login page.

Once they had logged in, they were requested to agree on a consent form mentioning the preservation of the anonymity of the data (see the *Consent and privacy* section below for details). Thirdly, the participants were taken to a questionnaire regarding personality, gender, highest level of education, and whether they were native English speakers or not (all the experiments were written in English). The questions regarding personality come from a piece of work by Gosling and Rentfrow [Gosling, Rentfrow, and Swann (2003)] and were used to estimate the five general personality traits. The questionnaire page is reported in Figure A.2.

Once the questionnaire submitted, the participants have access to the detailed instructions on the judgment process (see Figure A.3). After this step, they were taken to a waiting room until 6 participants had arrived at this step. At this point, they started the series of 30 games which appeared 3 at a time,



Questionnaire

Before we proceed to the instruction page and you start the game, we require you to answer a few questions. It is important that you trustfully answer the questions. This will help us to properly assess your results.

Here are a number of personality traits that may or may not apply to you. Please write rate each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

I see myself as ...

	Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
Extraverted, enthousiastic	○	○	○	○	○	○	○
Critical, quarrelsome	○	○	○	○	○	○	○
Reliable, self-disciplined	○	○	○	○	○	○	○
Anxious, easily upset	○	○	○	○	○	○	○
Open to new experiences, complex	○	○	○	○	○	○	○
Reserved, quiet	○	○	○	○	○	○	○
Sympathetic, warm	○	○	○	○	○	○	○
Disorganized, careless	○	○	○	○	○	○	○
Calm, emotionally stable	○	○	○	○	○	○	○
Conventional, uncreative	○	○	○	○	○	○	○

Are you a? ○ Man ○ Woman ○ don't want to answer

Are you a native English speaker? ○ Yes ○ No

What is the highest grade of school you have completed? ○ Elementary School ○ High school ○ College ○ Graduate

Submit

Fig. A.2 Questionnaire form.

with one lone round where they had to make judgments alone and two social rounds where the provided judgment being aware of judgments from others. An instance of the lone round is given in Figure A.4 for the *counting game* while a social round is shown in Figure A.5. In the *gauging game*, the question was replaced by "*What percentage of the following color do you see in the image ?*". For this type of game, a sample of the color to be gauged for each picture was displayed between the question and the 3 pictures. Accuracy of all judgments were converted to a monetary bonus to encourage participants to improve their judgment at each round.

At the end of the 30 games, the participants had access to a debrief page where was given the final score and the corresponding bonus. They could also provide feedback in a text box. They had to provide their email address if they wanted to obtain the bonus (see Figure A.6).

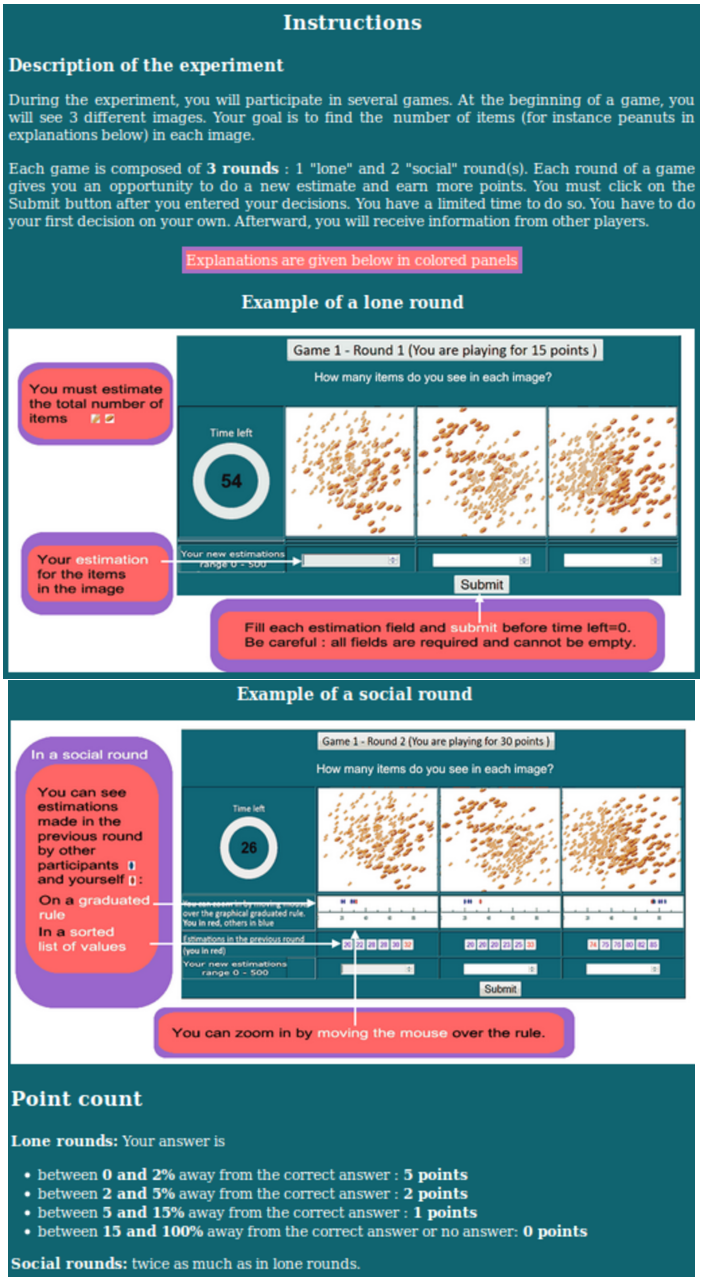


Fig. A.3 Instruction page.

(A)

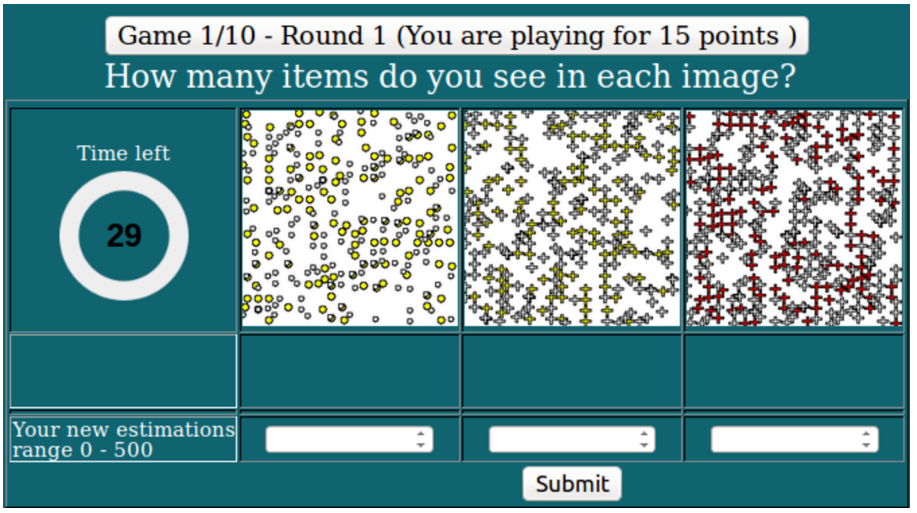


Fig. A.4 Interface of the first round for the *counting game*, played alone.

A.3 Consent and privacy

Before starting the experiment, participants had to agree electronically on a consent form mentioning the preservation of the anonymity of the data :

> *Hello! Thank you for participating in this experiment. You will be making estimations about features in images. The closer your answers are to the correct answer, the higher reward you will receive. Your answers will be used for research on personality and behavior in groups. We will keep complete anonymity of participants at all time. If you consent you will first be taken to a questionnaire. Then, you will get to a detailed instruction page you should read over before starting the game. Do you understand and consent to the terms of the experiment explained above ? If so click on* I agree *below.*

In this way, participants were aware that the data collected from their participation were to be used for research on personality and behavior in groups. IP addresses were collected. Email addresses were asked. Email addresses were only used to send participants bonuses via Paypal® according to their score in the experiments. IP addresses were used solely to obtain the country

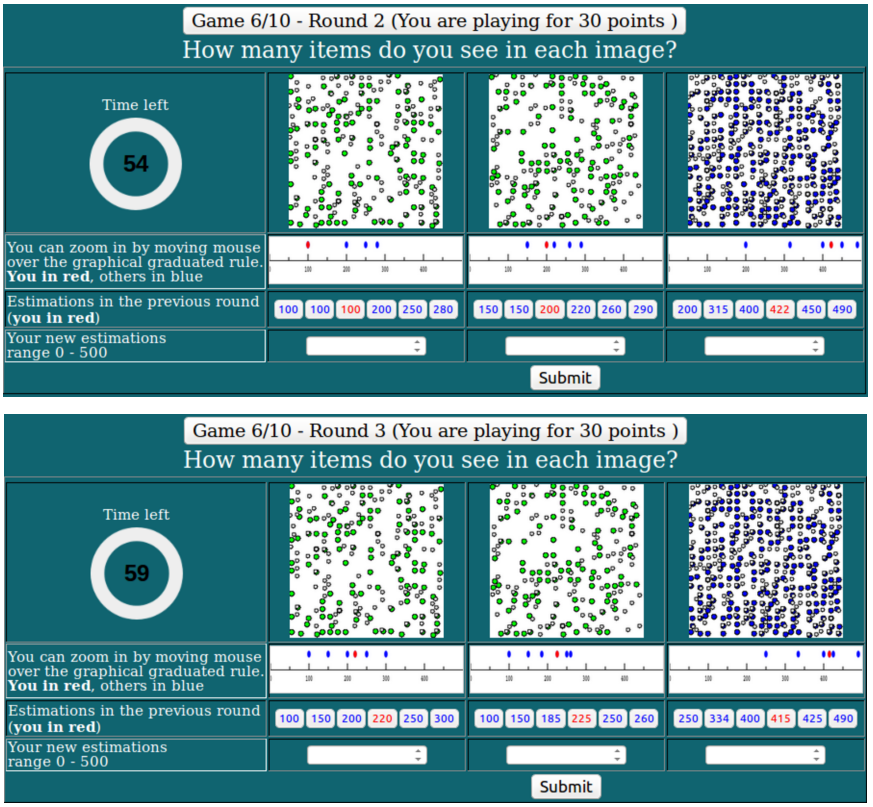


Fig. A.5 Interface of the second and third rounds for the *counting game*, social rounds

of origin of the participants. Behaviors were analyzed anonymously. Information collected on the participants were not used in any other way than the one presented in the manuscript and were not distributed to any third party. Personality, gender and country of origin presented no correlation with influenceability or any other quantity reported in the manuscript. The age of participants was not collected. Only adults are allowed to carry out microtasks on the CrowdFlower platform : CrowdFlower terms and conditions include : "you are at least 18 years of age". The experiment was declared to the Belgian Privacy Commission (https://www.privacycommission.be/) as requested by law. The French INSERM IRB read the consent procedure and confirmed that their approval was not required for this study since the data were analyzed anonymously.

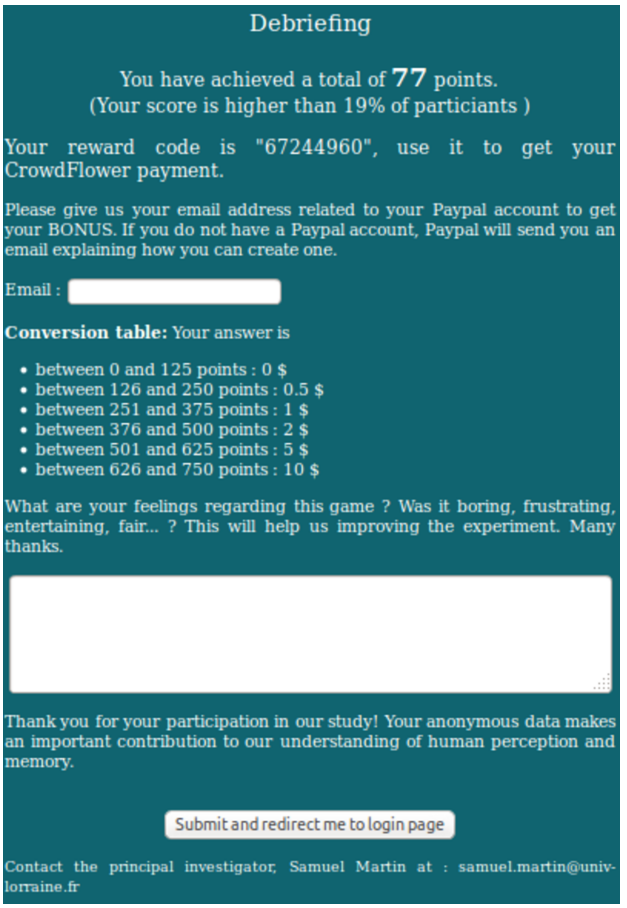


Fig. A.6 Debrief page.

Appendix B

Prediction accuracy : additional comments

B.1 Confidence intervals

Figure B.1 displays error bars for 95% confidence interval of the RMSEs. This figure reveals that the two methods depending on training set size do not perform significantly better than the consensus model with one couple of typical influenceabilities, even for large training set sizes. This is an argument to favor the model in which the whole population has a unique couple of influenceabilities $(\alpha(1), \alpha(2))$.

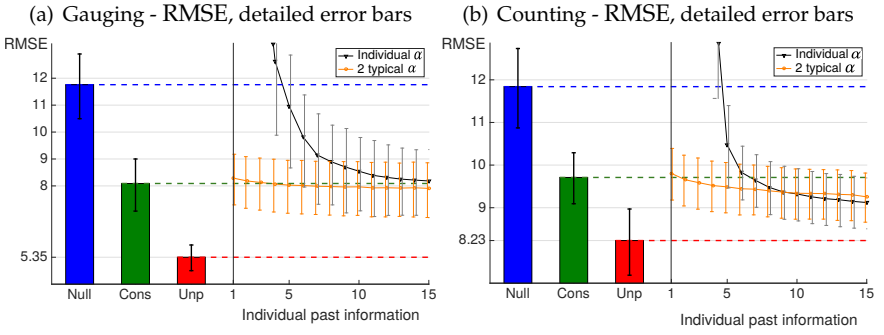


Fig. B.1 Root mean square error (RMSE) of the predictions with detailed error bars for the final round.

B.2 Mean Absolute Errors

Measuring prediction accuracy in terms of MAEs may appear more intuitive for comparing prediction methods. Figure B.2. assesses the models using an absolute linear scale, where the errors are deliberately unscaled for the *counting game*. The prediction methods rank equally when measured in terms of MAE or RMSE. Notice that, due to nonlinear relation between RMSE and MAE, on this alternative scale, the consensus models errors are now closer from the null model than from the unpredictability error. For comparison, recall that for the *gauging game*, the judgments range between 0 and 100 while they range between 0 and 500 for the *counting game*.

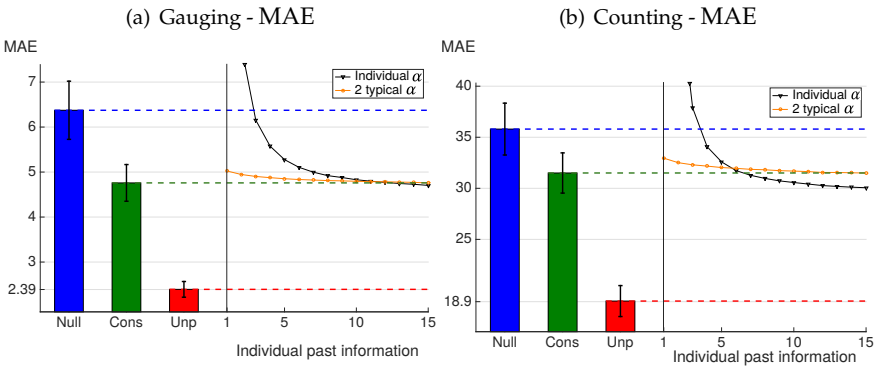


Fig. B.2 Mean absolute error (MAE) of the predictions (unscaled) for the final round.

B.3 Second round predictions

The prediction procedure based on the consensus model (4.6-CMOD) is applied to predict the second round judgments. Crossvalidation allows to assess the accuracy of the model. Results are presented in Figure B.3. These results are qualitatively equivalent to the prediction errors for the third round as shown in Figure B.3, although the second round predictions lead to a lower RMSEs, as expected since they correspond to shorter-term predictions.

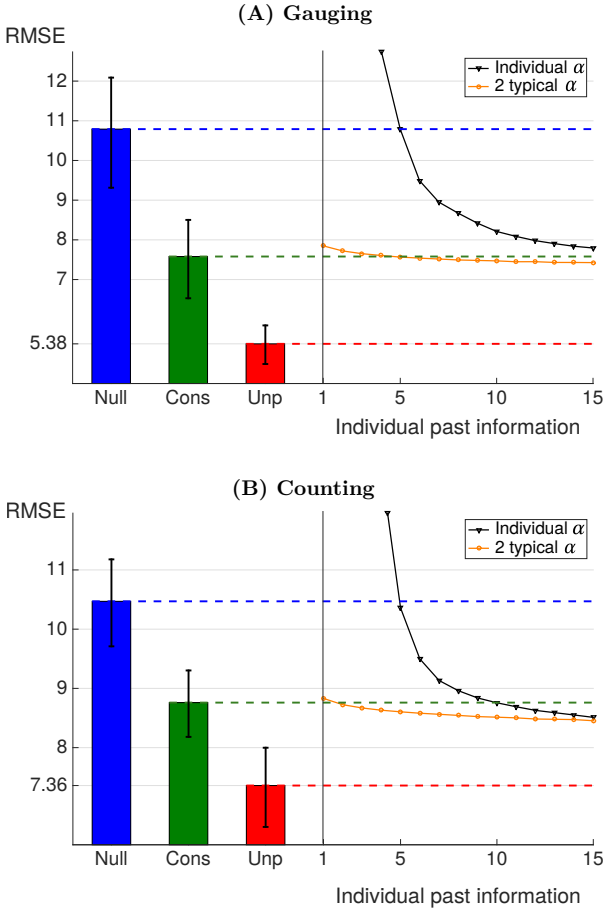


Fig. B.3 Root mean square error (RMSE) of the predictions for the *second* round. The RMSEs are obtained from crossvalidation. The bar chart displays crossvalidation errors for models that does not depend on the training set size. All RMSEs were obtained on validation games.

Appendix C

Political issues : dictionaries and questionnaires

In Chapter 5, we presented how we extracted a political dictionary from the Twitter posts of political elites during the election period. Table 5.2 presented a sample of the Canadian dictionary for the 2015th election. We present here the dictionaries for the two other contexts (New-Zealand 2014 and Quebec 2014). We then provide for each election the series of 30 questions proposed in the VPL survey which served as a validation for the textual estimates.

C.1 Unsupervised dictionaries

Table C.1 provides the most extreme bigrams in the context of New-Zealand 2014. The highest weights are associated to environmental issues or green slogans (*clean-river*,*vote-green*). Lowest weights include mainly references towards the *KiwiSaver* scheme (a savings scheme for retirement). In August 2014, John Key (leader of the National Party) proposed a new grant for first-home buyers in the KiwiSavers program [Gulliver (2014)], which fast became one of the major election issue. Weights are consistent with party positions of Figure 5.10(b).

Finally, the dictionary for Quebec 2014 is presented at Table C.2. As for Canada 2015 and New-Zealand 2014, the table displays the highest and lowest weighted bigrams. The greatest weights are direct references towards either the Option Nationale party *on-slogan* or the Québec Solidaire *solidair-action* party. This is consistent with Figure 5.11 since the parties are displayed closely in the political space. On the right side of Table C.2, bigrams are either taglines or major issues

Highest Weights				Lowest Weights			
Bigram		Weight	Popularity	Bigram		Weight	Popularity
lovenz	green20	9.92	-11.54	kiwisav	homestart	-5.74	-8.75
vote	green	7.73	-8.03	homestart	grant	-5.74	-8.75
green20	lovenz	7.66	-8.71	buyer	five	-4.98	-7.97
clean	river	6.45	-7.89	grant	help	-4.98	-7.97
vote	lovenz	6.16	-7.29	year	working4nz	-4.98	-7.97

Table C.1 Subsample of the political dictionary for New-Zealand 2014. The table displays the 5 lowest and highest weighted bigrams in the stemmed form.

Highest Weights				Lowest Weights			
Bigram		Weight	Popularity	Bigram		Weight	Popularity
action	rejoign	7.33	-10.58	pet	tour	-3.95	-5.92
solidair	action	7.33	-10.58	election2014	qc2014	-3.67	-2.72
rejoign	visit	7.33	-10.58	pme	plq	-3.61	-5.94
visuel	th	7.18	-10.75	souffl	plan	-3.5	-6.35
on	slogan	7.18	-10.75	pme	besoin	-3.5	-6.35

Table C.2 Subsample of the political dictionary for Quebec 2014. The table displays the 5 lowest and highest weighted bigrams in the stemmed form.

of the Quebec Liberal Party. As an example, the QLP leader Philippe Couillard quoted « *Les PME ont besoin de souffler, et le plan du Parti libéral du Québec leur donnera l'oxygène nécessaire* », which may explain the presence of bigrams such as *souffl-plan* and *pme-plq*.

C.2 Vox Pop Labs' survey

The selection of 30 issues per election follows a two-part process: a team of political scientists first analyzes the party platforms of each party and a larger set of issues is selected from parties' public statements. They pre-define each party's position regarding the first set of issues. In a second step, the political parties are contacted and provided with the opportunity to review these pre-defined positions, and a subset of 30 issues are selected. The selection

criterion includes mainly the question's ability of differentiation between the respondents[1].

The questions or assertions – presented in the following tables – differ from one election to another and are encoded on Likert scales ranging from 1 (strongly disagree) to 5 (strongly agree).

	Canada 2015
Issue ID	Question
1	How much should Canada do to reduce its greenhouse gas emissions?
2	No new oil pipelines should be built in Canada.
3	How much power should unions have?
4	The most effective way to create jobs in Canada is to lower taxes.
5	How many new immigrants should Canada admit?
6	How much should be done to accommodate religious minorities in Canada?
7	Terminally ill patients should be able to end their own lives with medical assistance.
8	Abortions should be allowed in all cases, regardless of the reason.
9	Possession of marijuana should be a criminal offence.
10	Longer prison sentences are the best way to prevent crime.
11	Handguns should be banned in Canada.
12	To what extent should law enforcement be able to monitor the online activity of Canadians?
13	Government workers should not be allowed to strike.
14	How much tax should corporations pay?
15	How much should wealthier people pay in taxes?
16	How supportive should Canada be of Israel?
17	How much should Canada spend on foreign aid?
18	How involved should the Canadian military be in the fight against ISIS?
19	How much of a role should the private sector have in health care?
20	Illicit drug users should have access to safe injection sites.
21	How much should the government do to make amends for past treatment of Aboriginal Peoples in Canada?
22	Aboriginal Peoples in Canada should have more control over their ancestral territory.
23	Canada should introduce a publicly funded childcare program.
24	Canada's budget should be balanced no matter what.
25	Quebec should be formally recognized as a nation in the Constitution.
26	Quebec should become an independent state.
27	Canada should end its ties to the monarchy.
28	The Canadian government should put a price on carbon.
29	The Senate should be abolished.
30	Only those who speak both English and French should be appointed to the Supreme Court.

[1] This is achieved by comparing factor loadings on pre-election surveys

	New-Zealand 2014
Issue ID	Question
1	How much should the government spend on rebuilding Christchurch?
2	How much should government do to reduce the gap between rich and poor?
3	When businesses make a lot of money everyone benefits, even the poor.
4	When there is an economic problem, government spending usually makes it worse.
5	How much government assistance should welfare recipients receive?
6	The government should restore free tertiary education.
7	Linking educators' salaries to student results is more effective than cutting class sizes.
8	The government should use National Standards to monitor and improve education.
9	How much influence should trade unions have in the workplace?
10	How much should the minimum wage be?
11	The age when people receive New Zealand Superannuation should be gradually increased.
12	How much funding should the Department of Conservation receive?
13	The government should not allow further fracking.
14	How much of a role should the private sector have in New Zealand's health care system?
15	Up to what age should the government fund GP visits for children?
16	The government should build affordable housing for Kiwis to buy.
17	How much state housing should the government provide?
18	The government should prevent foreign ownership of New Zealand farms.
19	How many immigrants should New Zealand admit?
20	The anti-smacking law should be repealed.
21	Sentences for repeat offenders should be stiffer than they are now.
22	How much support should there be for the Maori language?
23	How much of a role should the Treaty of Waitangi have in New Zealand law?
24	How much control should Maori have over their own affairs?
25	Abortion up to twenty weeks should not require medical approval.
26	Marijuana should be legalized.
27	How old should Kiwis be before they are allowed to purchase alcohol?
28	How much tax should corporations pay?
29	How much should wealthier people pay in taxes?
30	Landlords should pay more tax on the sale of their rental properties.

Quebec 2014	
Issue ID	Question
1	When businesses make a lot of money everyone benefits, even the poor
2	Teachers' salaries should be tied to their performance, rather than seniority
3	Education should be free from kindergarten through university
4	Environmental regulations should be stricter, even if it means consumers pay higher prices.
5	There should be a ban on oil extraction in Quebec
6	How much should wealthier people pay in taxes?
7	How much tax should corporations pay?
8	First Nations should have more control over their ancestral territory
9	The crucifix should continue to be displayed in the National Assembly
10	Quebec should force the federal government to hand over full control of communications and culture
11	Quebec's budget should be balanced no matter what
12	Services for the elderly should be funded entirely by the government
13	Drivers should be taxed more in order to pay for public transportation services
14	Unions have too much influence in Quebec
15	A referendum on Quebec independence should be held within the next five years
16	How much of a role should the private sector have in health care?
17	The government should pay the full cost of prescription drugs
18	Government employees should not be permitted to wear religious symbols or clothing while at work
19	The government should privatize Hydro-Quebec
20	The government should add tolls to more roads and bridges
21	Elected officials should be allowed to cover their faces for religious reasons
22	Quebec should become an independent state
23	French should be the mandatory working language in small and medium-sized businesses
24	How many immigrants should Quebec admit?
25	How much should be done to accommodate religious minorities in Quebec?
26	Patients suffering from incurable illness should be able to end their own lives with the assistance of a doctor
27	How much money should welfare recipients receive?
28	Students from French schools should no longer have access to English CEGEPs
29	How much effort should the government put into protecting the French language in Quebec?
30	How much money should the government spend on arts and culture?

Appendix D

QR codes

Election or Country	URL	QR code
Canada 2015	https://cvandekerckh.github.io/mythesis/canada.html	
New-Zealand 2014	https://cvandekerckh.github.io/mythesis/new-zealand.html	
Quebec 2014	https://cvandekerckh.github.io/mythesis/quebec.html	
Belgium - Wallonia	https://cvandekerckh.github.io/mythesis/wallonie.html	
Belgium - Flemish	https://cvandekerckh.github.io/mythesis/flandre.html