UNIVERSITY OF PISA

Department of Computer Science

MSc in Data Science and Business Informatics

MASTERS' THESIS

MODELING PEER-PRESSURE: A DATA-DRIVEN ESTIMATE OF OPEN-MINDEDNESS FROM HIGH-ORDER ONLINE POLITICAL DISCUSSIONS

RELATORI

Prof. Giulio ROSSETTI

Dr. Valentina PANSANELLA

CANDIDATO

Giulia FERRO

ANNO ACCADEMICO 2022-23

Abstract

One of the main research areas of Social Network Analysis is Opinion Dynamics, which studies how individuals form, share and update their opinions in social networks. The necessity of going beyond pairwise interactions modeled by ordinary graphs has led to the development of higher-order structures that allow modeling non-linear multiagent interactions. This change requires also the adaptation of Opinion Dynamic models to the new configuration, which is one of the goals of this thesis. In this work, we focus on confidence bound models, where each individual interacts with their peers only if the difference between their opinions is smaller than a given threshold, which can be interpreted as their level of open-mindedness. This study adopts a data-driven approach to estimate the user's open-mindedness following a Deffuant-like procedure applied on real networks which are defined from online political discussions in the U.S. In fact, the possibility to collect new information regarding users' interactions due to the increment of online social network platforms usage, allows to provide empirical validation of the models. The goal is not only to estimate the users' open-mindedness in real networks, but also to understand if and how the two different underlying structures influence the results of the procedures.

Contents

1 Introduction

One of the most debated and analyzed phenomena on online social networks is the tendency to observe political polarization [1, 2, 3, 4], i.e., the divergence of political attitudes to ideological extremes not aiming at reaching any form of synthesis. Indeed, social media blurred the boundaries of communication and democratized content diffusion. However, the fact that we can potentially interact with much different information does not mean that we can engage with and be influenced by opposing or even mildly different stances [5]. Humans, indeed, are far from being perfectly rational individuals and are affected by a series of cognitive biases[1] in the process of forming their belief system. Among these, it is worth mentioning confirmation bias [6, 7], the human tendency to ignore content that counters their prior beliefs by a) choosing to interact with like-minded individuals [8] b) ignoring interactions with the "opposite faction" [9]. Unfortunately, by connecting with like-minded individuals and ignoring contrasting views, users can easily compromise their open-mindedness [10], exacerbating and polarizing their views even more and, ultimately, giving rise to the formation and persistence of echo chambers [11].

One of the most exploited approaches for understanding the effects of different biases on public opinions - especially political opinions - is through mathematical models of opinion formation [12], where parameters incorpo-

[1]Cognitive biases cheat sheet: `shorturl.at/afjpM`, last visited June 2022.

rate psychological factors (e.g., cognitive biases) affecting individual opinion evolution. The so-called *bounded confidence models* constitute a broad family of models where agents are influenced only by peers having an opinion sufficiently close to theirs. This characteristic is justified by sociological theories such as **homophily**. This phenomenon can foster the formation of echo chambers on social networking sites, where people with similar ideologies only interact with each other. This tendency describes the creation of bonds mainly between similar individuals [8] - but can also represent, for example, a lack of understanding, conflicts of interest or close-mindedness [9].

Opinion dynamics models provide insight into the social phenomena that have an impact on information diffusion, opinion formation and group dynamics, but despite their versatility, they lack empirical validation [13].

However, thanks to the advent of the Internet - and with the rise of social media - an increasing part of human interactions leave a massive digital footprint that can be exploited to study the dynamics of opinion formation and diffusion.

Following such reasoning, since different models or parameter values can predict different, even opposite, effects of biases on opinions [14], there is a crucial need for empirical works to study and quantify socio-psychological and external drivers of the dynamics.

Although online social networks offer a vast opportunity to retrieve people's opinions, friendships, interactions, and discussions, more quantitative analysis of real data still needs to be done. Empirical approaches are claimed to be the next necessary step by many researchers in the field.

The current lack of data-driven approaches that validate models in real settings has led Pansanella et al.[15] to develop a data-driven time-aware methodology [16] that estimates users' open-mindedness, starting from users' interactions represented as networks.

However, in many online contexts (e.g., Reddit), people mainly participate in group discussions, which could be better captured by exploiting higher-order structures. The intrinsic nature of network connectivity can not explicitly go beyond dyadic patterns, i.e., coupling between pairs of nodes. Such pairwise constraints must be considered when investigating social group behavior and its complex dynamics. Hypernetwork science is the new, cutting-edge line of research that addresses higher-order structures for representing and analyzing complex (social) systems.

This study builds upon the methodology employed by Pansanella et al. and aims to achieve two primary objectives. Firstly, it aims to estimate the distribution of open-mindedness among users participating in diverse discussions that share the commonality of contentious topics. Secondly, it aims to compare the outcomes generated by network-based modeling and higher-order structure-based modeling, to assess whether the latter can provide a more comprehensive understanding of users' tendencies in group dynamics.

The remainder of this thesis is organized as follows:

- Chapter 2 contains a summary of the literature necessary to understand the methodology presented and the analyses performed. In particular, there is an introduction to the field of social network analysis, with a focus on graph theory and higher-order structures and then there is a summary of the opinion dynamics models employed in the present work.

- Chapter 3 presents the analyzed datasets and the process implemented to collect them, which can also be found in [4]. The Chapter continues with the definition and description of networks and hypergraphs built from the collected interactions.

- Chapter 4 explains the open-mindedness estimation methodology on networks (already introduced in [15]) and its extension to hypergraphs.

- Chapter 5 discusses the results of the analysis. It is mainly divided into two parts, the first one shows the results of the pairwise graph for each dataset. The second deals with the insights gained from the procedure implemented on the hypergraph.

- Chapter 6 discusses similarities and differences between the results obtained from the two different methodologies.

- Chapter 7 discusses the main contributions of the work, its limitations and the future possible applications.

2 State of the art

This Chapter provides a theoretical foundation for understanding the data structures and methodology presented in subsequent Chapters. Specifically, it first introduces Social Network Analysis and subsequently discusses pairwise interaction graphs, their characteristics and properties. The Chapter then explores the transition from ordinary pairwise graphs to hypergraphs, which are used to model higher-order interactions. Furthermore, the Chapter introduces the Opinion Dynamics field, with a particular focus on the Deffuant-Weisbuch model, which is essential for the analysis in this thesis. Finally, the Chapter concludes with a discussion of the data-driven analysis approach employed in this thesis.

2.1 Social Network Analysis

The concept of connectedness is one of the characterizing aspects of modern society. The systems in which connectivity is a key feature are represented by networks, which are structures used to describe the pattern of interconnections among a set of elements [17]. Given this broad definition, it is evident that there are several domains of applications where the network structure can be used to model the complexity characterizing the systems. Some of the systems that can be modeled as networks are:

- *Technological and infrastructural systems.* Infrastructural systems, like computer networks or databases, are defined as organized network that facilitates information accessibility and sharing among different devices.

- *Information systems.* In Information systems, the elements of the network are pieces of information related one to another or linked for other specific reasons. The most known information network is the World Wide Web (WWW) [18].

- *Transportation systems.* Networks are also used to implement transportation systems, like highways or railways, in the most effective and efficient way to facilitate the movement of people or the transportation of goods.

- *Natural world systems.* In the biochemical field, networks are used to represent ecosystems, food web, biochemical cycles and biological systems like protein interaction and metabolic networks. The connectivity of these systems models the flows of energy, nutrients or other elements that influence the functionality of the ecosystems.

- *Social systems.* In social systems, network structure represents patterns of interactions, communications and relationships between individuals, communities, organizations, societies and institutions.

Computational social science [19] is an interdisciplinary field that applies computational methods to the study of social systems and human behavior. It involves computer-based tools and techniques to collect, process, and analyze large-scale data sets, such as social media data or survey data, in order to gain insights into social phenomena. The field draws on a range of disciplines, including computer science, statistics, sociology, psychology, and economics, and has become progressively important as an increasing number of social interactions and behaviors are mediated by digital technologies. Computational social science encompasses a wide range of research topics, including social network analysis, text analysis, machine learning, and agent-based modeling, and has the potential to generate new insights into social phenomena, inform policy decisions, and develop new technologies that can improve human well-being. Computational social science and social network analysis are closely linked, as both fields rely on the use of computer-based methods to analyze and understand social phenomena. Social network analysis, in particular, is a branch of computational social science that focuses on the study of social networks and their properties, such as the patterns of connections between individuals or groups. By using computational tools to analyze large amounts of data from social networks, researchers can gain insights into a wide range of social phenomena, such as the spread of information, the formation of social groups, and the dynamics of social influence. In this way, computational social science and social network analysis provide complementary perspectives on the study of social systems, with the former providing the computational tools and methods, and the latter providing the theoretical framework for understanding social networks and their properties. Social network analysis employs tools from network science to analyze and understand the properties of social systems represented by networks. The representation of complex systems as networks enhance their comprehensibility by defining the interactions between their various components. A key challenge in analyzing complex systems is the difficulty of inferring collective behavior from individual components. Networks offer a useful approach in this context, allowing individual components to be represented as nodes and their relationships as links. Additionally, networks retain additional information on the components and capture a range of characteristics that would otherwise be difficult to comprehend. By using networks, researchers can better understand the behavior of complex systems and derive insights that would otherwise be unattainable. The study of networks relies on the use of graphs as the mathematical and data structure. In a graph, vertices represent nodes while edges represent links. To analyze networks, a fundamental understanding of graph theory is required, which will be presented in the subsequent sections of this thesis.

2.1.1 Graph theory

In computer science and mathematics, a graph is a data structure that consists of a set of nodes, also known as vertices, and a set of edges that connect pairs of nodes. Visually speaking, the nodes are represented by circles, while the edges are the lines that link them. The formal definition of a graph G consists in a pair $G = (V, E)$, where $V = \{v_1, v_2, v_3, .., v_N\}$ is the non-empty set of nodes of G, while $E = \{(u, v)|u, v \in V\}$ is a set of edges or links between the vertices. A graph is undirected if the tuples representing the edges of the graph are not ordered: the edges are not oriented and they connect two nodes in a symmetric relation. In this case, given an edge $e = (u, v) = (v, u)$, the vertices u and v are said to be adjacent. On the other hand, if the tuples are ordered pairs, i.e. $e_1 = (u, v) \neq e_2 = (v, u)$, the graph is directed. In general, if two nodes are linked with each other they are called neighbors. In addition to the orientation, another edge feature is its weight. In a weighted graph, each edge has a value assigned to it, which aims to represent the relationship between the nodes it connects with a quantitative measure. On the opposite side, all edges in an unweighted graph have the same weight, usually set equal to 1. One of the key properties of a node i is its degree, defined like k_i, which represents the number of its neighbors for undirected graphs. For direct graphs, it is necessary to distinguish between in-degree k_i^{in}, which is the number of incoming edges of the node i and out-degree k_i^{out} which is the number of outgoing edges of the node i. The total degree of the node is given by the sum of the in-degree and out-degree, $k_i = k_i^{in} + k_i^{out}$.

From the degrees, the total number of links L can be computed. In undirected graphs, the L is given by the sum of the nodes' degrees divided by 2 (eq. 1), since every link is counted twice because $e = (u, v)$ contributes both to k_u and k_v. In direct graphs, the total number of links L is simply the sum of all the in-degrees and out-degrees (eq. 2).

$$L = \frac{1}{2} \sum_{i=0}^{N} k_i \tag{1}$$

$$L = \sum_{i=0}^{N} k_i^{in} = \sum_{i=0}^{N} k_i^{out} \tag{2}$$

The average node degree $\langle k \rangle$ can be computed as:

$$\langle k \rangle = \frac{1}{N} \sum_{i=0}^{N} k_i = \frac{2L}{N} \tag{3}$$

In equation 3, L is multiplied by 2 because each edge contributes to the degree of 2 nodes.

Referring to undirected graphs, the maximum number of edges is given by

$$L_{max} = \binom{N}{2} = \frac{N(N-1)}{2} \tag{4}$$

where N is the total number of nodes. A graph where $L = L_{max}$ is called a complete graph, where each node is linked to every other node and the average node degree is equal to $< k >= N - 1$.

However, real networks (especially social networks) are very sparse [20] and the number of edges can take a wider range of values which is usually much lower than the maximum, $L << L_{max}$.

Real-world networks, such as social networks, technological networks, and biological networks, are often sparse, meaning that only a small fraction of possible connections between nodes are actually present in the network. This sparsity is a fundamental characteristic of real-world networks and has important implications for their structure and dynamics. However, despite their sparsity, real-world networks also exhibit another striking feature, namely scale-free behavior, which means that the distribution of node degrees follows a power law, with a few nodes having many connections and most nodes having only a few.

The Barabasi-Albert (BA) model [21] is a widely used mathematical model for generating scale-free networks. In this model, new nodes are added to the network one at a time, and each new node is connected to a fixed number of existing nodes, chosen with a probability proportional to their degree, according to the "reach-get-reacher" principle [22], which is formalized as *preferential attachment* in this context. As a result, nodes with a high degree tend to attract more links, leading to the emergence of a few highly connected nodes, known as hubs. The BA model has been shown to capture many of the structural properties observed in real-world networks, such as the small-world property and the scale-free degree distribution. Moreover, it has been used to study a wide range of phenomena, including epidemic spreading, information diffusion, and opinion formation, among others.

Connectivity

The concept of network connectedness can be seen from a double perspective, by focusing on the structural aspects, or by reasoning about the behavioral context and understanding how the single agents' interactions have consequences on the rest of the system. The following paragraph focuses on the first point of view and explores some of the measures to analyze it.

The first step is to define some fundamental concepts and give some definitions. In graph theory, a path is a succession of nodes where each consecutive pair in the sequence is linked by an edge. A node j is reachable from another node j if there is a path between them. With these notions, a graph is considered connected if every node i is reachable from any other node $j \neq i$, otherwise, it is divided into connected components.

A component in a network is a set of nodes that are connected to each other by a path, such that any two nodes within the component can be reached from each other through that path. The component cannot accommodate any additional nodes without breaking this property. If by removing an edge, the graph becomes disconnected, the removed edge is called a bridge. In other words, the elimination of a bridge cause the increase in the number of connected components w.r.t. the original graph.

In addition to the components analysis of the network, it's useful to study the regions that are more densely connected, to better understand the structure of the graph.

The **density** of a graph G is a measure given by the ratio between the total number of edges L and the maximum number of edges that G can have (eq. 4) and it is defined as follow:

$$D = \frac{L}{L_{max}} \tag{5}$$

A high density indicates that a large proportion of possible edges in the network are present, while a low density suggests that the network is more sparse and less connected. While density refers to the overall level of connectivity in a network, to measure the extent to which nodes cluster together locally we can use the **clustering coefficient**. This measure represents the degree to which nodes in a network are interconnected and reflects the presence of local clusters or communities within the network. A high clustering coefficient indicates that the nodes in the network tend to form tightly interconnected groups, while a low clustering coefficient suggests that the nodes are more sparsely connected and do not form many clusters. Given a node i, its clustering coefficient C_i is defined as

$$C_i = \frac{2L_i}{k_i(k_i - 1)} \tag{6}$$

where k_i is the node's degree and L_i is the number of links between i's neighbours. The local clustering coefficient takes values between 0 and 1, where the maximum value is obtained when the node i and its neighbors form a clique.

The average clustering coefficient indicates the degree of clustering of the entire graph. It can be seen as the probability that two neighbors of a randomly selected node are linked. The global clustering coefficient C is computed as the average of the local clustering coefficients of nodes $i = [1, .., N]$:

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{7}$$

The global clustering coefficient provides a measure of the prevalence of triangles, or closed three-node subgraphs, in a network. A triangle is formed when three nodes are connected to each other by three edges, forming a closed loop. In the local clustering coefficient formula, L_i can be seen as the number of triangles that the node i participates in, as each link between two neighbors of the node i closes a triangle. Hence the degree of the network's clustering can be also captured by the global clustering coefficient, defined by Watts and Strogatz [23] as:

$$C_\delta = \frac{3 \times \text{Number Of Triangles}}{\text{Number Of All Triplets}} \tag{8}$$

where a connected triplet is an ordered set of three nodes ABC such that A connects to B and B connects to C.

Given the sparsity of real networks, these measures of clustering coefficients are used to study the degree to which nodes tend to create tightly knit groups and the role of the triadic closure mechanism for their formation.

The principle of triadic closure states that given three nodes A, B, and C, if A and B are linked, and B and C are linked, there is a higher probability that B and C are also linked, creating a triangle. This concept was first introduced by George Simmel and developed by Mark Granovetter [24]. In social network theory, this recalls the idea that if two people have a friend in common, then there is a higher chance that they will meet and become friends themselves at some point or at least establish a weak connection.

Network dynamics

The limitations of static networks to accurately capture the evolution of systems where nodes and/or edges undergo changes over time have led to a growing interest in dynamical networks, which are complex systems consisting of interconnected elements that exhibit time-varying behaviors and have been extensively investigated across various disciplines such as physics, biology, engineering, and social sciences [25].

There are several advantages from using temporal networks instead of static networks, such as: (i) capturing dynamic behavior, (ii) studying real-world data, (iii) performing robustness analysis, (iv) and control and optimization. Since in these structures both nodes and links can appear and disappear over time, in social network analysis the volatility of people's interactions can be better modeled using dynamic networks with respect to static graphs. Usually, a dynamic network is modeled with two kinds of time-aware analytical processes: stream graphs [[26],[27]] and snapshot graphs [25].

In the latter formalization, the evolution of the network is partitioned into a series of snapshot graphs that represent the state of the network at a specific time, or the aggregation of the interactions during an established time window. Formally, this model is defined as $G = (G_1, G_2, \ldots, G_n)$ where each G_i is the i-th snapshot of the graph. The snapshot $G_i = (V_i, E_i)$ consists in the set of nodes V_i and the set of edges E_i active at time i [28].

The information of the snapshots can be aggregated in four different ways [28]:

- *Single Snapshot.* A snapshot graph represents exactly the nodes and the edges present in the selected instant of time. In some applications of this configuration, there can be information loss regarding the connection pattern between consecutive snapshots, especially if the selected instants are too far apart.

- *Disjoint window.* Here longer periods of time are allowed to accumulate information. Differently from the previous technique, no information is discarded: when a window ends, the next one begins immediately. Works well when it's not important to maintain continuity.

- *Sliding windows.* Is similar to the Disjoint window but in this case, the end of a time window is allowed to overlap with the consecutive one. That is, the next window starts before the previous one ends. Works well when it is important to maintain continuity.

- *Cumulative Windows.* Similar to sliding windows, but here we fix the beginning of each window at the beginning of the observation period. Information can only accumulate: edge information is never discarded, no matter how long ago it was first generated. Each window includes the information of all previous windows. Works well when the effect of an edge never expires, even after the edge has not been active for a long time.

In this thesis, we aggregated data using the disjoint windows technique, with each window corresponding to a semester of interactions.

2.1.2 From graphs to hypergraphs

Oftentimes, real contexts are too complex to be modeled with networks. One of the limiting aspects is that networks describe interactions in a pairwise fashion, while in complex systems the connections are not always established between pairs of entities but may rather be at a collective level [29]. A lot of social mechanisms, e.g. peer pressure, are led by multiple simultaneous interactions and not pairwise ones. To understand the transition from pairwise to a more complex interaction structure, it is needed a new set of definitions.

We define an interaction as a set $I = [p_0, p_1, ..., p_{k-1}]$ containing an arbitrary number k of basic elements of the system under study, which we indicate as nodes or vertices. The order, or dimension, of the interaction that includes k node is $k-1$: therefore, an interaction involving only a single node (node interacting with itself) is a 0-order interaction, an interaction between 2 nodes has order 1, one among 3 nodes has order 2, and so on. Interactions with $k \geq 2$, i.e. involving at least 3 nodes, are considered *higher-order interactions*. In plain terms, low-order systems are those in which only self or pair-wise interactions take place (like edges in a graph), while higher-order systems (HOrSs, from now on) display interactions in groups of more than two elements. As a result, an interacting system $(V, \mathcal{I})$ is composed by a set of nodes V that are involved in a collection of interactions $\mathcal{I} = I_0, ..., I_n$. The configuration with *higher-order interactions* can be described with the notion of hypergraphs which in mathematics is a generalization of the concept of ordinary graph exposed before.

An **hypergraph** $H = (V, E)$ is defined as a set of nodes V and a set of hyperedges E, which is a subset of the power-set of the nodes $\mathcal{P}(X) \setminus \emptyset$. In this case, a hyperedge can connect any number of nodes, while an edge of the ordinary graph can link only two nodes. The cardinality of the nodes set V is the order of the hypergraph, while the total number of hyperedges represents its size. The size of a single hyperedge is the number of nodes that are connected by it. If all the hyperedges have the same size k, the hypergraph is called *k-uniform*, which means that a 2-uniform hypergraph represents an ordinary graph where the connections are pairwise. If a hyperedge has cardinality 1 is called a singleton and it means that it is incident to only one node. The degree of a node d_i is the number of hyperedges that include it. If all the nodes have the same degree, the hypergraph is called *d-regular*. Similarly as in graphs, in this case, two or more nodes are called *neighbors* if they are included in the same hyperedges. In this structure, the node u's ego-network is the subgraph composed by the node u itself and all the hyperedges that include it.

2.2 Opinion Dynamics

An understanding of the individual and social mechanisms of opinions change is essential to explore the development of collective movements, trends and large-scale social phenomena.

Opinion Dynamics (OD) is a multidisciplinary research field where mathematical models and computational tools are used to understand the dynamic processes of the diffusion and evolution of opinions in social networks. Such models generally consider a population of individuals (called agents) and simulate the interactions between them numerically - whenever possible, they compute the final state analytically. Such processes are normally governed by rules - often even very simple equations - developed according to empirically observed sociological behaviors, chosen by scientists in an attempt to reproduce patterns observed in the real world and provide a causal explanation of them.

Before digging into the details of the different models, we need to formalize some notions to understand how opinions and their dynamics can be mathematically represented. Three elements are essential and must be precisely defined in every model: opinions, interactions and time.

Opinions. In OD models, each agent i holds an opinion x_i, which can be discrete or continuous, based on the specific context that has to be studied. The set of all possible opinions is called opinion space X. The opinion of an agent i at time t is $x_i^{(t)}$. We denote the vector of the opinions of the whole population at time t as $x^{(t)} = [x_1^{(t)}, x_2^{(t)}, ..., x_N^{(t)}]$, where N is the size of the studied population.

In discrete OD models, the opinion variable can be binary, e.g., $x_i \in \{0, 1\}$ or $x_i \in \{-1, 1\}$ or any subset $X = \{x_1, x_2, ..., x_k\}$ with $x_i \in \mathbb{N}$, to represent multiple but finite choices. Some of the most well-known OD models that deal with discrete opinion variables are: the Voter model [30, 31], the Majority Rule model [32] and the Sznajd model [33]. On the other hand, in continuous models, opinions are real-valued variables, e.g. $x_i \in [0, 1]$ or $x_i \in [-1, 1]$, to express, for example, the level of agreement or disagreement with a specific view or the political leaning of an individual.

One of the key insights of network theory is that the underlying network structure can strongly affect dynamic processes that are mediated by the edges. In many classical OD models, a mean-field approach is adopted, i.e. the underlying network is modeled as a complete graph, where the possible interactions are $\frac{N(N-1)}{2}$, where N is the number of agents (nodes). While this choice can be suitable for some contexts, e.g., a small group of people discussing a topic, the assumption that every individual is connected to every other individual is quite strong and may rarely be applied to real contexts. In fact, people aggregate in social networks - offline and online. This limits the set of possible interactions among agents and makes them subject to network effects. While the mean-field approximation is still widely employed as a starting point, further analyses and extensions of established models often employ generative network models as an underlying structure. In these cases, the underlying network structure is a graph where nodes represent agents (i.e. individuals), while edges represent social connections and thus possible interactions between two agents: through such interactions agents may share and update their opinion.

2.2.1 Bounded Confidence Models: the Deffuant-Weisbuch Model on graphs

Bounded-confidence models are continuous models where agents interact with each other only when their opinions are "similar enough". Among them, there is the Deffuant-Weisbuch model [9]. This model starts by considering a population of N agents i with opinions that lie in a finite real interval $x_i(t) \in [0, 1]$. At each interval, two random agents (i, j) interact with each other. They are influenced by the interaction if the distance $|x_i - x_j|$ between their opinions is smaller than a threshold ϵ. Given x_i the opinion of the agent i, x_j the opinion of the agent j and that $|x_i - x_j| < \epsilon$, the rule for the opinion update is:

$$\begin{aligned} x_i(t+1) &= x_i + \mu(x_j - x_i) \\ x_j(t+1) &= x_j + \mu(x_i - x_j) \end{aligned} \tag{9}$$

Where $\mu \in [0, 0.5]$ is the convergence parameter.

The confidence bound ϵ, as introduced in Chapter 1, can be seen as a measure of the open-mindedness of the population. Low values of ϵ mean that the individuals are influenced only by agents that have a similar opinion, while high values of ϵ imply that the agents are more open to engage in discussions and to be influenced by people with dissimilar ideas.

The model simulations on complete graphs show that the qualitative dynamics depend on the value of ϵ, which influences the number of peaks of the final opinions distribution which decreases as a function of ϵ. On the other hand, the convergence time and the width of the final opinions distribution are impacted by the parameter μ and the number of agents N.

One of the limitations of this model, explicitly mentioned in [9], is that the parameter ϵ is considered constant for all the population. In fact, it's very unlikely that all the agents share the same level of open-mindedness (like for [34], [35], [31]).

2.2.2 Bounded Confidence Models: the Deffuant-Weisbuch Model on hypergraphs

The necessity of going beyond pairwise interaction modeled by ordinary graphs, satisfied by the definition of hypergraphs structure, requires also the adaptation of the Opinion Dynamic Models to the new configuration.

The dynamic rule of the Deffuant-Weisbuch model for pairwise interactions (eq. 9), states that at each time step the two agents' opinion is updated only if their opinion difference is less than the confidence bound.

If we change the underlying social structure of the model from a graph to a hypergraph, interacting pairs become interacting groups modeled by hyperedges.

In the HOID model [36] the dynamic rule determines that at each timestamp a random hyperedge e is selected and every node included in e updates its opinion with the following rule:

$$x_i(t+1) = \begin{cases} \overline{x_e} & \text{if } max_{j \in e} x_j(t) - min_{j \in e} x_j(t) \leq \epsilon \\ x_i(t) & \text{otherwise} \end{cases} \tag{10}$$

Where $\overline{x_e} = \frac{1}{|e|} \sum_{j \in e} x_j$ is the average opinion of the agents linked by the hyperedge e. This means that the opinion update happens only if the opinion differences between all the interacting neighbors are less than the given confidence bound ϵ. In other words, a context influences the agents' opinions only if it doesn't include users with opinions too distant with respect to the rest of the group, which, instead, precludes the possibility to reach consensus. This kind of group interaction is different from the one proposed in the Hegselman-Krause [37] model: the difference stands in the fact that in this higher-order model, the distances between all nodes within the hyperedge matter, while in the HK model, only the distance between a given node and its neighbor have an impact on the opinion change of the target node. As a consequence, a dissenter can block the interaction of all other agents in the hyperedge in the HOID model, a mechanism absent from the HK model.

Like with pairwise interactions, the confidence bound value is fundamental and has a great impact on the outcome of the model implementation. Clearly, a small confidence bound prevents a lot of group discussions to be influential, especially the more confrontational ones. Moreover, with a low confidence bound the probability of an interaction to be influential decays exponentially w.r.t. the size k of the hyperedge, due to the fact that larger groups of agents with very different opinions have a lower probability to reach consensus. While higher confidence bound levels bring to consensus, low values cause opinion fragmentation and polarization. The interesting insight drawn from [36] is that with the hypergraph configuration, there isn't the sharp transition to consensus like in [9]. When the hypergraph becomes dense, the smooth crossover that characterizes the transition to consensus still holds for hyperedges of $k = 6$, but even for smaller hyperedges the transition is still less sharp. In other words, a small confidence bound decrease doesn't cause an abrupt switch from consensus to opinion polarization, but a gradual decrease in the number of agents that share the same opinions. This mechanism of slow transition is considered more realistic to describe the stability of real social contexts.

On the other hand, it's also true that to overcome network opinion fragmentation is necessary to have a high confidence bound which allows unblocking larger hyperedges. Following this reasoning, it can be expected that the most influential contexts in a nonhomogeneous hypergraph (where the hyperedges are of different sizes) are the smaller hyperedges, but it is not the case unless they result to be significantly majoritarian. The dynamics on the heterogeneous hypergraph where the average hyperedges size is given by k, is similar to the k uniform hypergraph [36].

Similarly, in [38] a Hypergraph Bounded Confidence Model (HBCM) is defined. In this case, to generalize the notion of confidence bound to hyperedges, authors define a *discordance function* that maps a hyperedge and an opinion state to a real number and quantifies the level of disagreement among the nodes that are incident to a hyperedge, to determine whether or not these nodes update their opinions. Authors in [38] consider the following family of discordance functions:

$$d_\alpha(e, x) = (\frac{1}{|e| - 1})^\alpha \sum_{i \in e} (x_i - \overline{x}_e)^2 \tag{11}$$

which is parameterized by the scalar α, where $\overline{x}_e = \sum_{i \in e} / |e|$. If the discordance $d_\alpha(e, x(t)$ is less than the confidence bound c, the hyperedge e is concordant at time t. Otherwise, it is discordant. In the case of $\alpha = 1$ $d_1(e, x(t)$ is equal to the unbiased sample variance of the opinions of the nodes that are incident to e. The scaling parameter $\frac{1}{|e|-1}$ prevents advantaging hyperedges with few nodes over hyperedges with many nodes when there is an opinion update. In the model, at each time step an hyperedge $e \in E$ is selected at random according to some probability distribution (e.g. uniform distribution). If the discordance function is less than the confidence bound ϵ th nodes $i \in e$ update their opinions x_i to the mean opinion $\overline{x}_e$; otherwise, their opinions do not change. One way to think about this update is that nodes are "peer-pressured" into conforming to the mean opinion of the group when the overall discordance of the group is sufficiently small. More formally:

$$x_i(t+1) = \begin{cases} \overline{x_e} & \text{if } i \in e \text{and } d(e, x) \leq \epsilon \\ x_i(t) & \text{otherwise} \end{cases} \tag{12}$$

2.2.3 Bridging the gap between models and data

Although assumptions and simplifications are made in building the presented opinion dynamics models, they have proven very useful in explaining well-known phenomena in opinion formation. This literature on opinion dynamics models is wide, going from binary opinions and pair-wise interactions models and moving towards continuous opinions on time-evolving higher-order systems, trying to narrow the gap between the models and the real systems from a theoretical perspective. Conclusions from different models appear to be realistic and seem to explain some real-world phenomena in a plausible way. However, there is no agreement on which models and characteristics better represent social interactions and one of the reason is that there is a scarcity of research in this direction. Outputs from discrete models have been compared to patterns seen in the data, such as voter model and election output data. Recently, past voting records and the voter model were used to forecast election results in the US and UK [39]. In some cases distributions seen in real data from surveys [40, 41, 42], social experiments [43], or social media [44] are used to manually tune parameters or modify agent-based models. Together with epidemic and information spreading, modeling opinion dynamics is an archetypal example of how interconnected decisions of individuals lead to emergent collective behavior at the level of society. In epidemic modeling, microscopic rules and parameter values are informed by how biological pathogens work, while online communication datasets guide the development of information-spreading models. Data on opinions is more sparse, typically coming from detailed but small sociological surveys, or aggregated over population and time in polls and elections. This limitation has led to an explosion in opinion dynamics models without a counterpart in empirical validation. Despite online social networks offer such a huge opportunity to retrieve people's opinions, friendships, interactions and discussions, there is still a lack of quantitative analysis on real data and empirical approaches are claimed to be the next necessary step by many researchers in the field. Agent-based models of opinion dynamics have the advantage of including a causal mechanism that makes the models interpretable. However, they do not exploit the availability of data and parameter calibration is a manual and difficult task. In the last years, some researchers tried to tackle the opinion dynamics understanding problem using more empirical approaches. Some studies employ Bayesian learning techniques. Monti et al. in [45] proposed a learnable generalization of an opinion dynamics model [46] and tried to estimate the backfire effect and latitude of commitment of a political discussion on Reddit. This kind of approach maintains the causal interpretation possible while allowing for model selection and hypothesis testing on real data. In this study, the only observables considered are actions and interactions, while in [47] opinions are considered fully observable and estimation of parameters through maximum-a-posteriori is used to find the most influential nodes. The approach is applied to Twitter and Reddit datasets. In [48] an ad-hoc model of opinion dynamics is developed and then Bayesian inference is used to calibrate model parameters from real data. The model was further developed in [49] where each user is assigned a recurrent neural network to learn non-linearly from past timings and opinions. However, application to real data to validate models' conclusions is still very scarce and the lack of this kind of approach is one of the two major issues addressed in our work. Traditional data sources, like surveys and questionnaires, can be useful tools in which people can explicitly express their ideas, but they usually offer a small sample size and they don't allow to collect observations with a temporal continuity making it impossible to have an insight on the network dynamic. The availability of large-scale social information about people's ideology and their evolution over time after interactions with others allows to adopt a data-driven approach in this field that will be exploited in this thesis.

3 Datasets: controversial discussions on Reddit

Due to the data-driven nature of this thesis, before digging into the details of the performed analysis, this Chapter introduces the dataset used. In this Chapter, we will describe the data used for our research. We will explain the rationale behind the choice of data collection, how it was retrieved, and how it was labeled with political opinions. Additionally, we will discuss how temporal networks and hypernetworks were constructed from the data. This information is crucial for understanding the results obtained and for evaluating the reliability of our methods.

3.1 Data collection and ideology estimate

Political expression and participation have taken a new form with the development of online social network platforms where individuals have not only access to news and information but also actively share and get in touch with other users' opinions. The constant news update and the growth of people's expressive potential through social networks have increased opinions exchange. One of the most used platforms for opinion sharing and political discussions is Reddit. This social platform is a news and forums website organized in communities. Each community, called subreddit, is a forum-like space, which is dedicated to a particular topic or interest. Users can submit content such as links, images, and text posts to the subreddit, and other users can comment and vote on the content. Each subreddit has its own set of rules and moderators who enforce them. Subreddits can range from broad topics like news, politics, and science to niche interests like a particular TV show, hobby, or fandom.

The data was previously retrieved for a study on network echo-chamber [4] with the Pushshift API [50].

The goal in [4] was to gather information on political debates about United State politics to have an insight into the interactions between the two political parties, Republican and Democratic.

The collected posts cover an interval from January 2017 to July 2019 - the first two years and a half of Donald Trump's presidency, who swore in as the 45th President of the United State on the 20 of January, supported by the Republican party. Like all of the Trump Era, this period was characterized by even more heated political discussions where users had very polarized opinions. Another important political event was the mid-term elections held on 6th November 2018, where the Democratic party took control of the U.S. House of Representatives and the Republican party controlled the U.S. Senate.

To have an insight into different sociopolitical issues and to study how the interactions change w.r.t the discussion argument, the data was collected from several subreddit via Reddit List and divided into three datasets. The first dataset we analyzed focuses on **Gun Control** policy (Gun Control, henceforth). To gain a comprehensive picture of the discussions on this topic, posts, and comments are gathered from subreddits that are both against and in favor of gun legalization. The second dataset contains data from discussions on **Minorities** discrimination (Minorities henceforth), including both users with more conservative ideas and others that encourage gender, sexual and racial equality. While it is important to understand how users may discuss very specific and controversial topics, we are also interested in understanding how they engage in general political discussions and the results that these debates have on the evolution of their interactions and opinions. Therefore, the last dataset we analyzed, which we will refer to as **Politics** henceforth, includes posts and comments from political discussion and gives a general view of the U.S. sociopolitical sphere.

To label each user with an opinion retrieved from the text of their posts and comments, a BERT text classifier was trained to classify the political leaning of the Reddit posts. To do so, a ground truth dataset was created with Pro-Trump and Anti-Trump posts by selecting subreddits known to have very polarized positions (e.g. `r/The_Donald`, `r/FuckTheRight` and `r/EnoghtTrumSpam`). After the training step, the BERT model was applied to the three datasets (Gun Control, Minorities, and Politics) mentioned above, obtaining a range prediction between 0 and 1, where 1 represents Pro-Trump ideology and 0 Anti-Trump.

The assumption made in the present thesis is that the Pro-Trump ideology aligns with Republican political leaning, the Antri-Trump with Democrat, and the neutral with the Moderate. The derived users' political leaning scores $L_{u,s}$ are computed as the average value of the monthly post leaning for user u that took part in the discussions in each semester s

$$L_{u,s} = \frac{\sum_{p \in P_{u,s}} Prediction\_Score(p)}{|P_{u,s}|} \tag{13}$$

where $P_{u,s}$ is the set of posts shared by the user u in each semester s. The value $L_{u,s}$ ranges from 0 to 1, and it is discretized into three intervals: $L_{u,m} \leq 0.4$ for Democrats, $L_{u,m} \geq 0.6$ for Republicans and $0.4 \leq L_{u,m} \leq 0.6$ for Moderates. This third label is adopted to identify users with a political position that is not fully Democrat or Republican.

3.2 Network Definition

This section focuses on the organization and the characteristics of the graph built from the information collected in the previous section with the structure explained in section 2.1.1. After the data collection and pre-processing, the interaction networks were built from the Reddit discussions. The time range of two years and a half was discretized in five intervals of six months each. Each snapshot network G_s is constructed as an undirected weighted graph. Each graph is composed of a set of nodes V_s, where each node represents a user who participated in the discussion at time s. Users participated in the discussion by posting or commenting on other users' posts. An edge (u, v) is present between two users if they have commented on each other's posts. Additionally, each node in the network has an attribute that represents their political leaning, denoted as $L_{u,s}$, which is calculated as the average of the political leanings of each contribution made by user u to the discussion at time s (see 2.1.1). The weight of each edge (u, v) is represented by an integer number $w_s(u, v)$ that indicates the number of interactions (reciprocal comments) between users u and v. During this period, the number of users increases, and it is possible to have an insight into the evolution of the users' political leanings $L_{u,s}$ because more than the 60% of users are present across consecutive months, which is essential to be able to compute the open-mindedness level.

Gun control dataset						
Interval	Nodes	Edges	Avg Clustering Coeff.	Avg node degree	Components number	Density
01-07 2017	833	4044	0.1898	9.7095	8	0.01167
07-12 2017	847	3925	0.1687	9.268	6	0.010955
01-07 2018	1054	3942	0.1363	7.4800	4	0.0071
07-12 2018	985	3478	0.1090	7.0619	16	0.0072
01-07 2019	1046	3601	0.0904	6.8853	6	0.006589

Table 1: Snapshots graph properties for the Gun Control dataset for each time window

Minority dataset						
Interval	Nodes	Edges	Avg Clustering Coefficient	Avg node degree	Number of components	Density
01-07 2017	1040	3765	0.214	7.240	3	0.00697
07-12 2017	1004	3465	0.200	6.902	5	0.00688
01-07 2018	1170	3832	0.185	6.550	5	0.0056
07-12 2018	1113	3594	0.161	6.458	6	0.0058
01-07 2019	1126	3405	0.154	6.048	5	0.00538

Table 2: Snapshots graph properties for the Minority dataset for each time window

Politics dataset						
Interval	Nodes	Edges	Avg Clustering Coefficient	Avg node degree	Number of components	Density
01-07 2017	917	2525	0.165	5.507	4	0.0060
07-12 2017	746	1816	0.149	4.869	10	0.0065
01-07 2018	825	2179	0.138	5.282	5	0.0064
07-12 2018	775	1787	0.124	4.612	5	0.0059
01-07 2019	686	1411	0.098	4.114	11	0.0060

Table 3: Snapshots graph properties for the Politics dataset for each time window

In tables 1, 2, and 3 we summarized the networks' characteristics for each dataset during the five intervals. For Minorities and Gun Control, the number of nodes increases steadily over time and more users engage in discussions concerning these controversial topics during the Trump era. Important events may have influenced users' participation. For instance, in Gun Control data, the sudden increase of users in the third semester can be linked to the Parkland mass shooting that happened in February of 2018. All the datasets have more or less the same level of density, higher values can be found in the first interval of the Gun Control dataset, which shows also the highest values of average degree with respect to the other topics of discussion. All the graphs are divided into connected components. The following figures show the distribution of the political leaning of each dataset during the periods. While the Gun control (figure 1) and the Minority (figure 2) datasets are quite balanced during the periods - describing a good proportion of Republicans, Democrats and Moderate - the Politics dataset shows a very unbalanced situation (figure 3) where only a few of the users are Republicans.

Figure 1: **Gun Control dataset distribution of the users' political leanings.** Histograms of the users' political leanings distributions in the Gun control network divided by Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

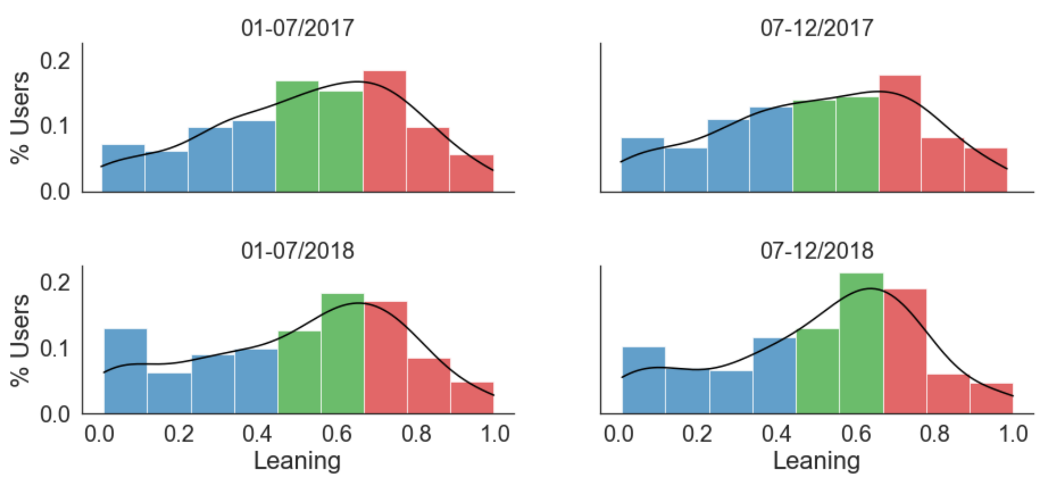


Figure 2: **Minoriy dataset distribution of the users' political leanings.** Histograms of the users' political leanings distributions in the Minority network divided by Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

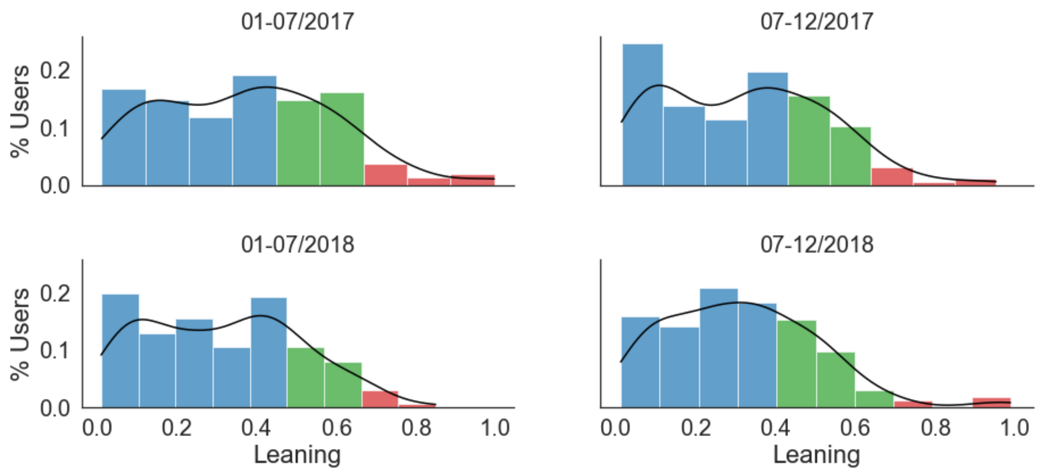


Figure 3: **Politics dataset distribution of the users' political leanings.** Histograms of the users' political leanings distributions in the Politics network divided by Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

3.3 Hypergraph Definition

The same datasets introduced in section 3, have also been to describe higher-order interactions whose characteristics were presented in section 2. The nodes and their opinion variables are obviously the same and also the temporal intervals of six months each, resulting in five hypergraphs snapshots. As we discussed in section 2.1.2, assuming that interactions take place only in a pairwise fashion is not always a good representation of reality, especially for online political discussions, where each post usually includes multiple comments and users interact in groups.

In a Reddit post, comments can have different levels of "indentation" based on the specific orientation of a user's response to another comment, creating smaller sub-group discussions. To translate this aspect with the hypergraph structure, the nodes are linked by a hyperedge if the users commented at the same level of indentation. Leveraging the original temporal networks, we infer the hypergraph structure by means of all the maximal cliques. In this way, the hyperedge becomes the context of interaction of the nodes that are included in it, highlighting the multiplicity of points of view that simultaneously participate in the discussion.

Not every hyperedge links the same number of nodes, so an insight into the distribution of the hyperedge dimension is shown in figure 4 respectively for Gun Control, Minorities, and Politics dataset. It can be seen that most of the hyperedges include few nodes meaning that the contexts of interactions analyzed are mostly small. The majority of nodes are included in less than 25 hyperedges for Gun Control and less than 10 for the other two datasets.

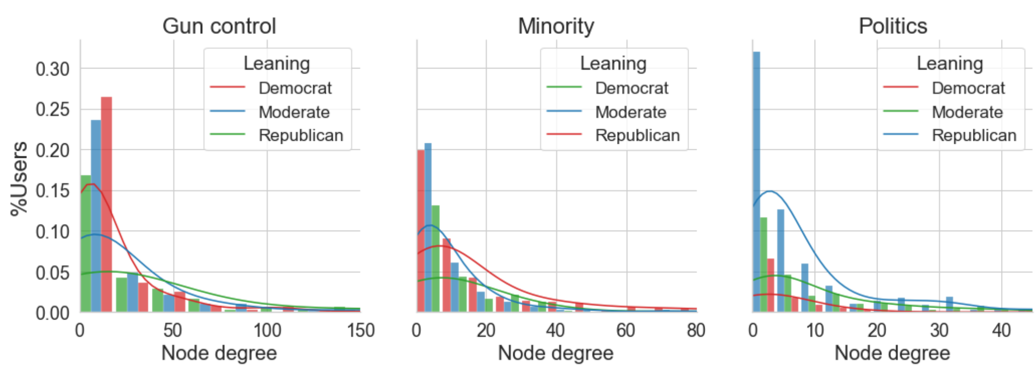


Figure 4: Node degree distributions for each dataset (Gun control, Minority and Politics) in the hypergraphs.

4 Methodology

This Chapter presents the methodology employed to estimate the open-mindedness of users. The first section delineates the implementation of the algorithm on ordinary graphs and the information that is collected. The second section expounds upon the adaptation of the same framework to a hypergraph configuration. The final section concentrates on interpreting the significance of the newly obtained measures.

4.1 Estimating open-mindedness on networks

The process of open-mindedness estimation on networks was first implemented in the study "Change my Mind: Data-Driven Estimate of Open-Mindedness from Political Discussions" by Pansanella et al. [15].

In Opinion Dynamics models, agents update their opinions after interacting with their neighbors according to simple mathematical rules. As we introduced in Chapter 2, in [9] agents average their opinion with the opinion of their interacting peer, randomly chosen from their neighbors' pool, if and only if their opinion distance is below a certain threshold, which we can interpret as the open-mindedness of the population. The hypothesis of the original model that open-mindedness is a characteristic trait of an entire population and not a characteristic that varies from individual to individual is strong and probably unrealistic. For this reason, in the present work, we assume a Deffuant-like process of opinion update (i.e., users averaging their opinions with the opinions of their interacting partners in a pairwise fashion) and provide a data-driven time-aware estimate of individual-level open-mindedness.

The Algorithm 1 shows the steps of the process that allows estimating the individual *confidence bound* $\widehat{CB}$, which represents the level of open-mindedness specific for each user. The necessary condition to estimate the level of open-mindedness is, of course, that the node u is present in two consecutive time intervals, i.e. the information about its current $x_u(t)$ and following $x_u(t+1)$ opinion value must be available. Once we checked this condition, all the opinions of the node u's neighbors $v \in N_{u,t}$ are ordered from the closest to the furthest by the opinion distance absolute value $d_{u,v} = |x_u(t) - x_v(t)|$ (Alg. 1 line 6). Respecting this ordering, we compute a final estimate $\widehat{x_u}(t+1)$ by iteratively averaging a temporary estimate $\widehat{x_u}(t+1)$ with the next interacting neighbor's opinion $x_v(t+1)$ (Alg. 1 line 13). The final estimated value $\widehat{x_u}(t+1)$ is the one that minimizes the error with respect to the real value $x_u(t+1)$ (Alg. 1 lines 15-22). The confidence bound is computed as the absolute value of the difference between the opinion of the node u, $x_u(t)$ and the opinion of the neighbor $x_v(t)$ used to compute the best estimation of $\widehat{x_u}(t+1)$ which is the one with the lowest error $e = |\widehat{x_u}(t+1) - x_u(t+1)|$.

Algorithm 1 Confidence bound estimation algorithm.

G_t = Weighted undirected interaction network at time t;
V_t = set of nodes at time t;
E_t = set of weighted edges at time t;
$x_u(t)$ = opinion of agent u at time t;
$d_{u,v} = |x_u(t) - x_v(t)|$ = opinion distance between $u, v \in V$ at time t;
$\widehat{CB}$ = estimated confidence bounds.

if $u \in V_{t+1}$ **then**
  Procedure to estimate $\widehat{x_u}(t+1)$ and $\widehat{CB}_u$
  $N_{u,t} = \{v|(u,v) \in E_t\}$; $|N_{u,t}| = n$
  $X_{N_{u,t}}[1...n]$ = Array of opinions of nodes $v \in N_{u,t}$
  if $N_{u,t} \neq \emptyset$ **then**
    Sort $X_{N_{u,t}}[1...n]$ *by* $d_{u,v}$ in ascending order.
    $\widehat{X_u}(t+1)[0...n]$ array of estimated opinions
    $\widehat{X_u}(t+1) = x_u(t)$
    $E = [0...n]$ array of estimation errors
    $E[i] = 1.0$
    for i=1; i=n; i++ **do**
      $x_v = X_{N_{u,t}}[i]$
      $\widehat{X_u}(t+1)[i] = \frac{\hat{X}_u(t+1)[i-1]+x_v}{2}$
      $E[i] = |\widehat{X_u}(t+1)[i] - x_u(t+1)|$
      $min_e = E[n]$
      for i=n; i=0; i– **do**
        $e = E[i]$
        if $e \leq min_e$ **then**
          $min_e = e$
          $j = i$
        end if
      end for
    end for
    $\widehat{x_u}(t+1) = \hat{X}_u(t+1)[j]$
    $\widehat{CB} = |x_u(t) - X_{N_{v,t}}[j]|$
  end if
end if

Measures

From the estimation procedure described in Algorithm 1 we collect the following measures:

- Observed opinion: real opinion of the node.
- Estimated opinion: mean between the node leaning score and the leaning score of the node which has the minimum error value.

- Error: difference between the estimated leaning score of each node and the observed one.

- Confidence bound $\widehat{CB}$: difference between the observed node opinion and the mean of the opinion node used to compute the best opinion estimation.

These measures are extracted after the application of the procedure on all the datasets, Gun Control, Minorities and Politics. Their distributions and interpretation are presented in Chapter 5.

4.2 Estimating open-mindedness on hypergraphs

In this section, we introduce the extension of Algorithm 1 to hypergraphs.

The first condition to proceed with the estimation remains the same: the node must be present in two consecutive intervals (Alg.2 line 1). In the network setting, the opinions of nodes adjacent to u in the time interval t (semester) were collected in the array $N_{u,t}$ (Alg. 1 line 3). In order to extend the algorithm 1 on hypergraphs, we need to modify this. Specifically, we must collect all the political leanings of the nodes included in the hyperedges of the star ego-network of node u. To accomplish this, we define an array of arrays $X_{C_{u,t}}$ (Alg. 2 line 5) where each nested array corresponds to a hyperedge in the star ego-network of u. The values contained in these arrays are the political leanings of all the neighbors in that hyperedge. The cardinality of the array $X_{C_{u,t}}$ is equal to the number of hyperedges in which node u appears.

To obtain the array $\overline{X}_{Cu,t}$, we compute the mean of the political leanings of all nodes present in each hyperedge that includes node u at time t (Alg. 2 line 8). This array corresponds to the $X_{N_{u,t}}$ array used in Algorithm 1 (line 3). This array will be used to estimate the opinion of the node at the consecutive interval $t+1$.

For each value in the array $\overline{X}_{C_{u,t}}$, we iteratively estimate the node u's opinion at time $t+1$ (Alg. 2 line 16) by averaging the current $\widehat{x_u}(t+1)$ with each element of $\overline{X}_{Cu,t}$.

The optimal estimated opinion $\widehat{x_u}(t+1)$ is determined by minimizing the discrepancy between the estimated value $\widehat{x_u}(t+1)$ and the observed opinion at the subsequent interval $x_u(t+1)$, as indicated in lines 20 to 22 of Algorithm 2. The confidence bound $\widehat{CB}$ is calculated as the absolute value of the difference between node u's opinion and the mean value of the hyperedge opinions that allowed the estimation with the least error, as shown in line 28 of Algorithm 2.

Algorithm 2 Confidence bound estimation algorithm for hypergraphs.

V_t = set of nodes at time t;
H_t = set of hyperedges at time t;
$C_{u,t}$ = set of hyperedges of the node u star ego-network;
$x_u(t)$ = opinion of agent u at time t;
$\widehat{CB}$ = estimated confidence bounds.

---

if $u \in V_{t+1}$ **then**
  Procedure to estimate $\widehat{x_u}(t+1)$ and $\widehat{CB}_u$
  $|C_{u,t}| = n$;
  if $C_{u,t} \neq \emptyset$ **then**
    $X_{C_{u,t}}[1...n]$ Array of opinions of the node u star ego-network
    $\overline{X}_{C_{u,t}}[1...n]$ Array for the average opinion for each hyperedge
    for i=0; i=n; i++ **do**
      $\overline{X}_{C_{u,t}}[i] = \frac{\sum C_{u,t}[i]}{|C_{u,t}[i]|}$
    end for
    $\widehat{X_u}(t+1)[0...n]$ Array of estimated opinions
    $\widehat{X_u}(t+1) = x_u(t)$
    $E = [0...n]$ Array of estimation errors
    $E[i] = 1.0$
    for i=1; i=n; i++ **do**
      $x_v = \overline{X}_{C_{u,t}}[i]$
      $\widehat{X_u}(t+1)[i] = \frac{\widehat{X}_u(t+1)[i-1]+x_v}{2}$
      $E[i] = |\widehat{X_u}(t+1)[i] - x_u(t+1)|$
      $min_e = E[n]$
      for i=n; i=0; i– **do**
        $e = E[i]$
        if $e \leq min_e$ **then**
          $min_e = e$
          $j = i$
        end if
      end for
    end for
    $\widehat{x_u}(t+1) = \hat{X}_u(t+1)[j]$
    $\widehat{CB} = |x_u(t) - X_{N_{v,t}}[j]|$
  end if
end if

The main difference between the two procedures is that in Algorithm 1 the estimated node u opinion is computed with the mean of the single neighbor opinion which allowed the lowest error, while in Algorithm 2 the same estimate is computed with the average opinion of the context of interaction of the node u that provided the lowest error.

In the first procedure, each interaction is pairwise between neighbors, in the second the interaction happens simultaneously between the nodes linked by the same hyperedge which is why the hyperedge is considered the interaction context.

Measures

In the following list, there is the information collected from the procedure 2 that we will describe in Chapter 5:

- Observed opinion: real opinion of the node.
- Estimated opinion: mean between the node leaning score and the mean of the leaning score of the hypergraph which has the minimum value of error.
- Error: difference between the estimated leaning score of each node and the observed one.
- Confidence bound $\widehat{CB}$: difference between the observed node opinion and the mean of the hypergraph used to compute the opinion estimation.
- Standard deviation: standard deviation of the array $\overline{X}_{C_{u,t}}$ which collects the means of the leaning score of the hyperlinks that includes the node u in the interval t. This value describes the context in which the node u participates: if the standard deviation of the array is high it means that the node engages in heterogeneous discussions, on the

contrary, if it is low the node participates in contexts where the agents share the same opinions.

- Dimension of the context: mean of the hyperedges dimensions in which the nodes participate in.

5 Results

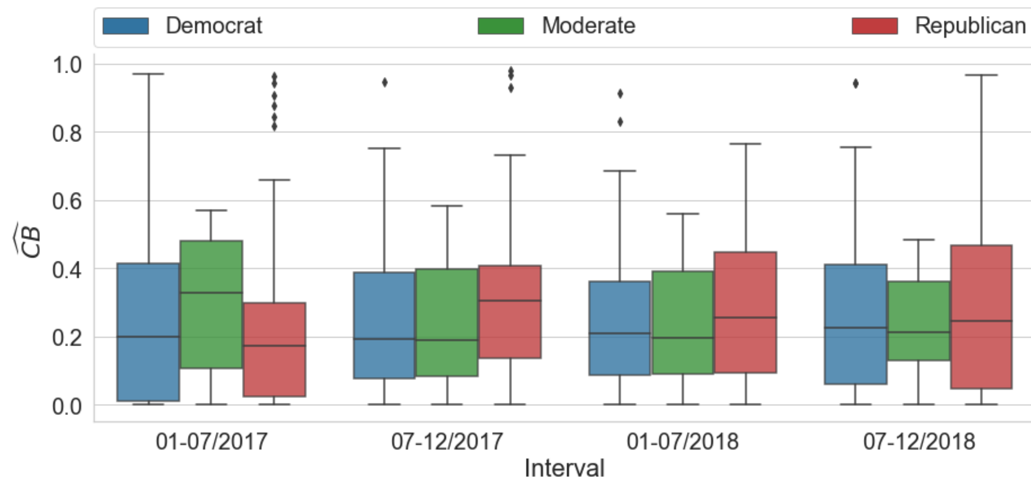
This Chapter presents and discusses the results obtained by implementing the methodology delineated in Chapter 4 on each dataset introduced in Chapter 3. The first section concentrates on the outcomes derived from the application of the open-mindedness estimation procedure to the pairwise interactions network. The second section exhibits the results obtained by applying the framework to the same data, modeled as hypergraphs.

5.1 Pairwise interactions

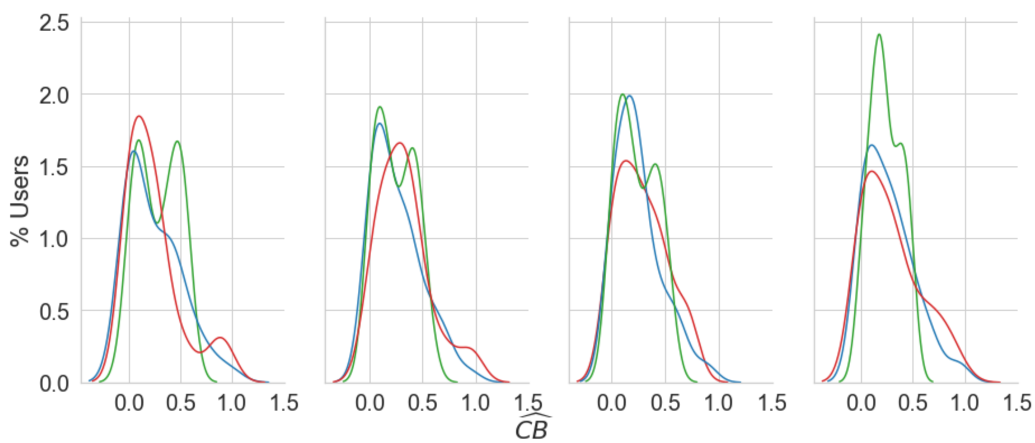
This section presents the outcomes of Algorithm 1 applied to the network data, as outlined in Chapter 4. We begin by discussing the results obtained from each of the three datasets - Gun control, Minorities, and Politics - individually. Subsequently, we investigate the comparison between the datasets, with the aim of comprehending how the topics of discussion impact users' open-mindedness.

5.1.1 Gun control

The first dataset analyzed is **Gun Control**. In this discussion, the distributions of the estimated open-mindedness $\widehat{CB}$ are quite stable over the four semesters (from 1/2017 to 12/2018) (see Figure 5).



(a)



(b)

Figure 5: **Gun control $\widehat{CB}$ distributions estimated on graph.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Gun control interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Considering the whole population, the estimated open-mindedness $\widehat{CB}$ presents a right-skewed distribution, with a median between 0.2 and 0.3, and 75% of the population with $\widehat{CB}$ under 0.4. The shape of the distributions is similar across the four semesters. Small differences can be seen in the second semester, which presents a narrower distribution but with a higher median. When we separately consider distributions for each discrete political leaning (Democrat, Moderate, and Republican), these distributions confirm that users in each population are rather close-minded (i.e. in each subpopulation, the third quartile is always below 0.5 and the median is always below 0.3). More in detail, while Democrats and Republicans have right-skewed distributions with longer tails that reach $\widehat{CB}$ of 1, all the Moderates have always a $\widehat{CB}$ lower than 0.8 and their distributions are more concentrated in a narrower range. This implies that most Moderates have the same level of close-mindedness, while the other two political leanings include users with more heterogeneous levels of open-mindedness. The Moderates distributions always present two peaks: one around a $\widehat{CB}$ of 0.1 and the other of 0.5, which become less pronounced in each interval. The first peak moves towards higher levels of open-mindedness. The distributions of the Democrats are the most consistent over time, their trend is more similar to the Republicans, especially in the last interval, and they both show higher variance with respect to the Moderate ones. Compared to the other groups, the Republicans are the most open-minded, especially in the second interval, which summed up discussions during the second half of 2017. During the considered period, two mass shootings took place[2], one of which was the Las Vegas shooting,

[2] https://en.wikipedia.org/wiki/Mass_shootings_in_the_United_States, last visited March 2023

that happened on the 1st of October 2017 and is currently the deadliest mass shooting committed by an individual in the U.S. Moreover, four mass shootings took place in the year from July 2017 to July 2018, which is included in the second and third snapshot networks.

The $\widehat{CB}$ distributions of the three political leanings are compared with a statistical test. Table 4 shows the results of the *Kolmogorov-Smirnov test* for each interval with the two sample configurations. The test quantifies the distance between the cumulative distribution functions. The null hypothesis of this test is that two independent samples are drawn from the same continuous distribution. This null hypothesis is rejected if the p-value is less than a specific significance level (normally 0.05).

		01-07 2017	07-12 2017	01-07 2018	07-12 2018
KS-stat	Dem vs. Rep	0.168	0.184	0.15	0.101
	Dem vs. Mod	0.258	0.131	0.113	0.177
	Rep vs. Mod	0.315	0.184	0.146	0.218
P-value	Dem vs. Rep	0.483	0.309	0.515	0.886
	Dem vs. Mod	0.131	0.73	0.792	0.331
	Rep vs. Mod	0.012	0.328	0.586	0.16

Table 4: **Kolmogorov-Smirnov test** on Gun Control pairwise graph to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

In table 4, we can see that the null hypothesis can be rejected only when comparing the Moderates' and the Republicans' distributions in the first semester (p-value $\leq$ 0.05). In the other comparisons the null hypothesis cannot be rejected, therefore we can assume that the confidence bound distributions samples are drawn from the same distribution.

Analyzing the confidence bound ($\widehat{CB}$) values at the user level provides valuable insights into the degree of change in open-mindedness over time. Figure 6 shows the standard deviation of the $\widehat{CB}$ values, with a higher deviation indicating greater variability in open-mindedness among users. Despite the presence of all political leanings, the standard deviations are generally low with right-skewed distributions, indicating that users' $\widehat{CB}$ values remain relatively stable over time. Notably, the Moderates exhibit the smallest change in their $\widehat{CB}$ values, suggesting they have a consistent open-minded level throughout the study period. On the other hand, the highest standard deviation levels are reached by the Republican subpopulation.

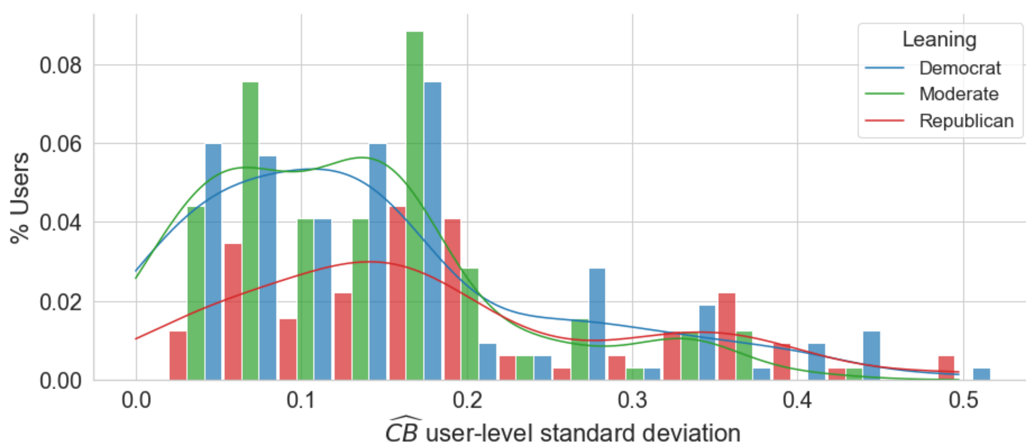
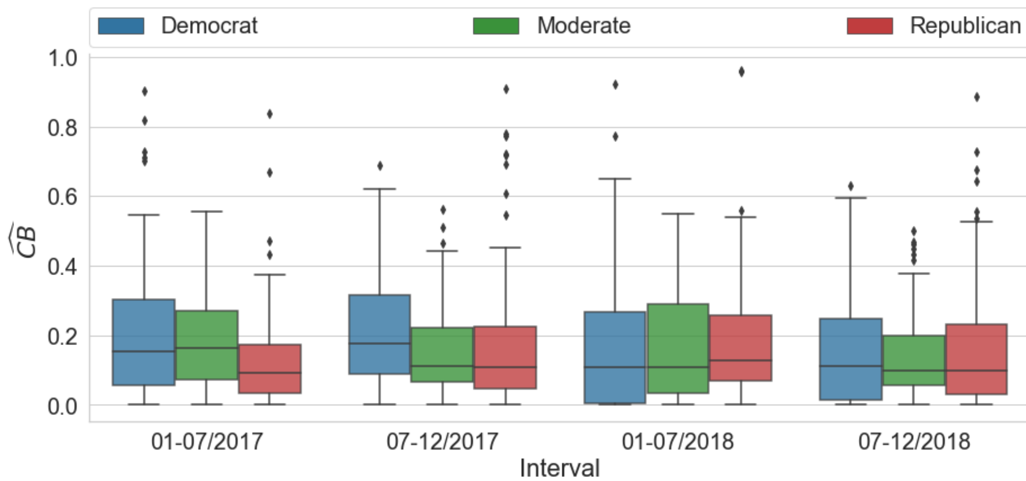


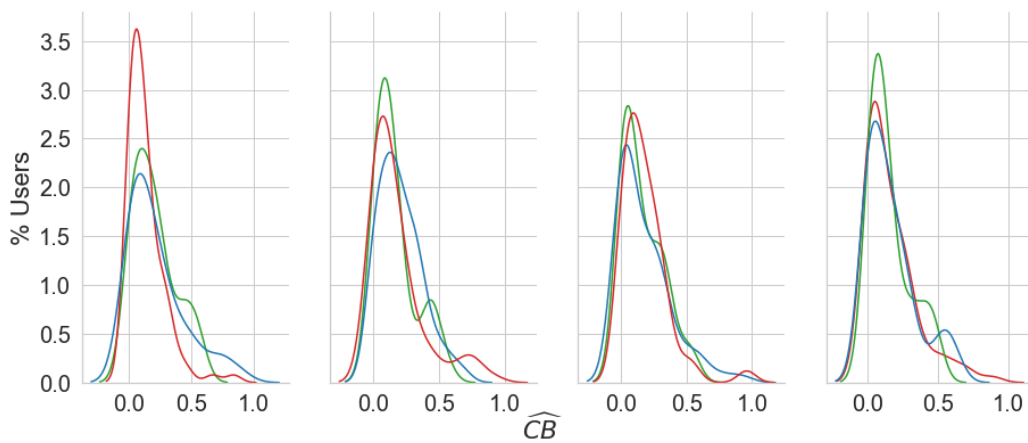
Figure 6: $\widehat{CB}$ user-level standard deviation distribution for the Gun control graph with pairwise interactions

5.1.2 Minorities

In this section, we present the analysis of the results obtained from the **Minorities** dataset. The estimated open-mindedness $\widehat{CB}$ distributions remain constant over time and are observed to be even more positively skewed than those of Gun Control. The median $\widehat{CB}$ is approximately 0.1, indicating that users are more closed-minded when discussing minorities discrimination, with respect to gun control, which suggests that the topic may be particularly polarising, in the long term, according to the Deffuant model [9]. The Republican and Democrat distributions (see Figure 7) continue to have longer tails, whereas the Moderate distribution shows a narrower range of values. During the last three semesters, the Republican distributions' tails reach a level of open-mindedness of 1, but remain positively skewed, so the majority of users remain steadfast in their positions, regardless of the interaction contexts. In the second interval, during the mid-term elections, the Democrat open-mindedness peak shifts slightly to the right, with the median $\widehat{CB}$ reaching nearly 0.2. Nonetheless, this value remains low and may lead to future opinion polarization in the population. The Moderate distributions show two peaks, as seen in the Gun control dataset, but with more users falling in the lowest level of $\widehat{CB}$.



(a)



(b)

Figure 7: **Minority $\widehat{CB}$ distributions estimated on graph.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Minority interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

We performed a statistical test to compare the $\widehat{CB}$ distributions for each discretized political leaning within each semester. Table 5 presents the results of the Kolmogorov-Smirnov test used to compare distributions. The values in the table indicate that the null hypothesis can only be rejected when comparing the Republican with the Moderate distribution in the first interval (p-value ≤ 0.05), implying that such samples are not drawn from the same distribution. In contrast, the null hypothesis cannot be rejected in all other cases.

		01-07 2017	07-12 2017	01-07 2018	07-12 2018
KS-stat	Dem vs. Rep	0.222	0.22	0.176	0.068
	Dem vs. Mod	0.098	0.215	0.133	0.149
	Rep vs. Mod	0.267	0.142	0.142	0.143
P-value	Dem vs. Rep	0.052	0.04	0.126	0.987
	Dem vs. Mod	0.93	0.106	0.628	0.472
	Rep vs. Mod	0.018	0.477	0.467	0.431

Table 5: **Kolmogorov-Smirnov test** on Minority pairwise graph to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

Regarding the confidence bound user level standard deviation (figure 8), the distribution shows that the majority of users don't have a drastic change in their open-mindedness levels over this two years period. As already observed for the Gun control dataset (figure 6), most users have a standard deviation lower than 0.2.

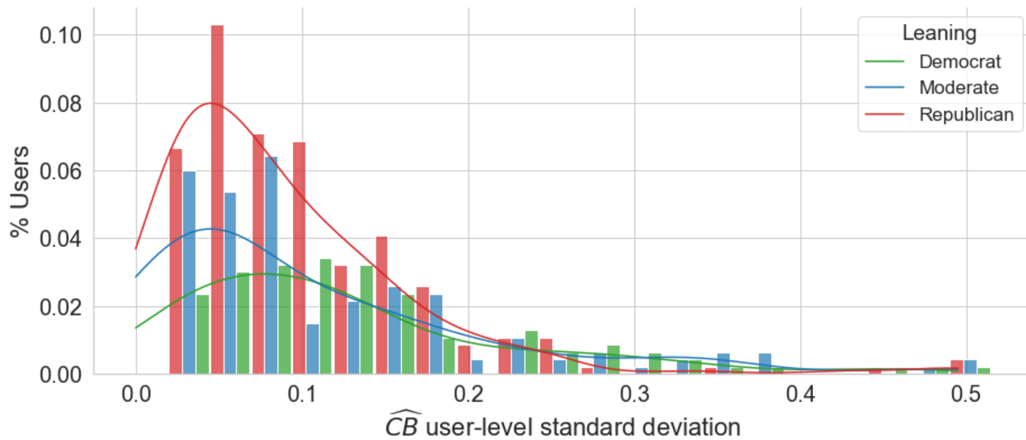


Figure 8: $\widehat{CB}$ user-level standard deviation distribution for the Minority graph with pairwise interactions

5.1.3 Politics

In this paragraph, we present the results relative to the Politics dataset. Our analyses showed that users tend to be close-minded even when discussing general socio-political topics. The confidence bound $\widehat{CB}$ distributions are highly positively skewed, with the majority of the values concentrated below the 0.2 threshold, meaning that during the periods, users tend to maintain their initial polarization. The most similar distributions are the Democrats and the Moderate, like for the Gun control dataset, which includes very close-minded users that in this case are even more close-minded with respect to the other datasets.

The table 6 shows the results of the Kolmogorov-Smirnov test.

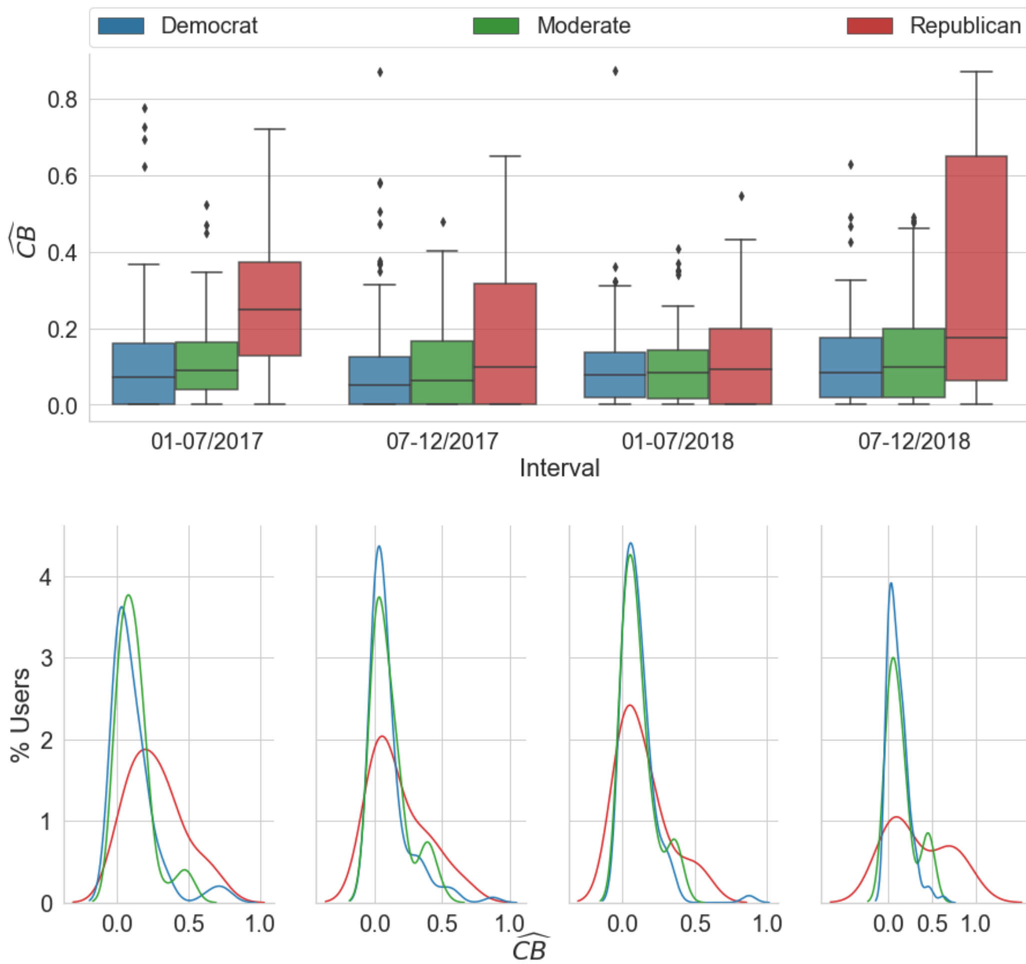


Figure 9: **Politics $\widehat{CB}$ distributions estimated on graph.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Politics interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Looking at the Kolmogorov-Smirnov test values in table 6, we can see that the values for the Dem vs. Rep and Rep vs. Mod pairs are higher than those for the Dem vs. Mod pair in all time intervals, indicating that the former pairs are less similar than the latter. In particular, the Dem vs.

Rep and Rep vs. Mod pairs have KS-stat values above 0.4, which suggests a relatively large difference between the two sample distributions. Regarding the p-values, we observe that in most cases the null hypothesis cannot be rejected, i.e. the p-value is higher than 0.05, suggesting that we cannot say that such sample distributions are statistically different. The only statistically significant difference emerges between Democrats and Republicans and between Republicans and Moderates in the first time interval (Jan-July 2017). Overall, the KS statistical test results suggest that there are some differences between the sample distributions of political leanings, with the Democrats and Republicans showing the largest differences. However, the significance of these differences varies across time intervals. The high diversity of the Republicans distribution is influenced by the unbalanced political leaning distribution: as observed in figure 3 there are only a few Republicans in the Politics dataset.

		01-07 2017	07-12 2017	01-07 2018	07-12 2018
KS-stat	Dem vs. Rep	0.458	0.252	0.231	0.4
	Dem vs. Mod	0.21	0.152	0.087	0.135
	Rep vs. Mod	0.454	0.226	0.216	0.4
P-value	Dem vs. Rep	0.0	0.317	0.458	0.076
	Dem vs. Mod	0.109	0.402	0.947	0.584
	Rep vs. Mod	0.001	0.535	0.612	0.115

Table 6: **Kolmogorov-Smirnov test** on Politics pairwise graph to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

Consistent with the previous datasets, the standard deviation demonstrates that most users do not experience significant changes in their open-mindedness over the observed time span (figure 10).

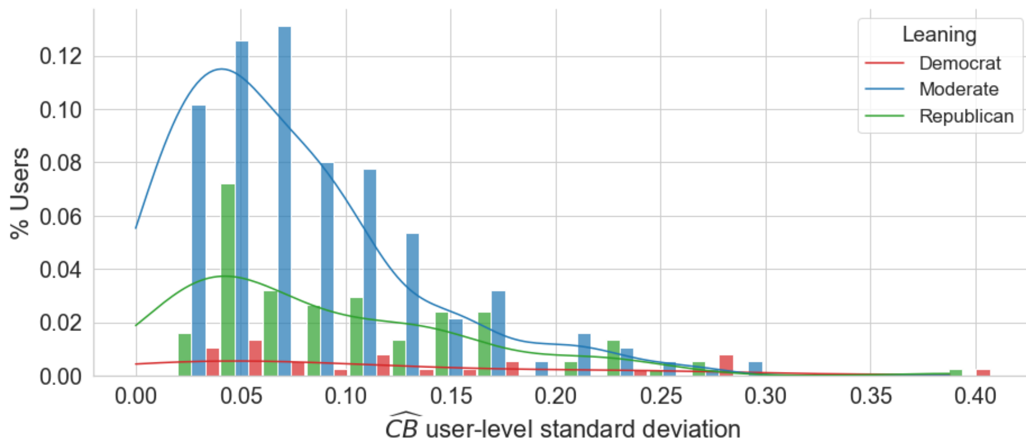


Figure 10: Politics dataset $\widehat{CB}$ user-level standard deviation for graph with pairwise interactions

5.1.4 Dataset Comparison

The previous sections focused on describing the estimated open-mindedness distributions in the three analyzed datasets. In particular, we highlighted the differences and similarities across different discrete political leanings. In the following paragraph, we instead compare the results and provide insights into how the three topics of discussion impact the users' level of open-mindedness.

In figures 11, 12, 13 we displayed the average and standard deviation of the estimated open-mindedness both for the entire population (left) and divided by political leanings (right). As we can see from the population-level distributions, in all datasets open-mindedness is considerably low, with average values below 0.2 in the Minorities and Politics datasets and below 0.3

in the Gun Control dataset. Concerning the temporal evolution, the open-mindedness is reasonably constant over the considered period, displaying just a small variation from one semester to the following. In the Minority discussion, the average level of open-mindedness decreases moving toward 2018. Considering the same values broken down by political leanings we can see that Republicans are the most open-minded when discussing about Gun Control and Politics, while they are initially more close-minded (on average) when discussing about Minorities, but their behavior tends to conform to the one of Democrats and Moderates towards the end, showing an opposite trend with respect to the other two political groups, which present a decrease in the level of open-mindedness over time. Moderates and Democrats show consistent behavior over the three datasets: almost no variations when discussing about Gun Control and Politics and a slight decrease when discussing about Minorities. Moreover, across the three datasets Moderates are always the most close-minded group.

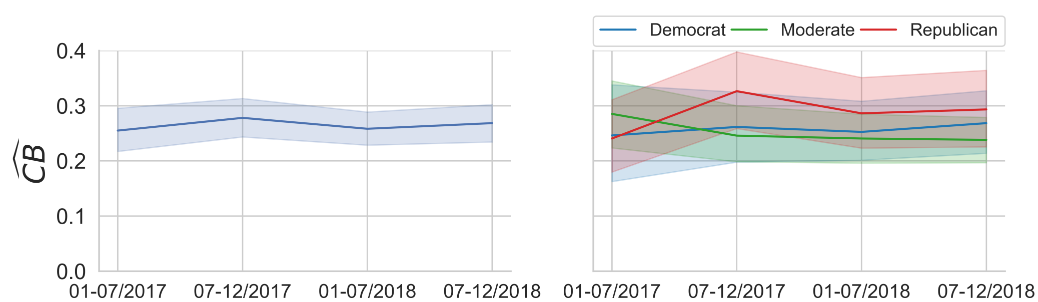


Figure 11: Confidence bound $\widehat{CB}$ mean trend of the Gun control graph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

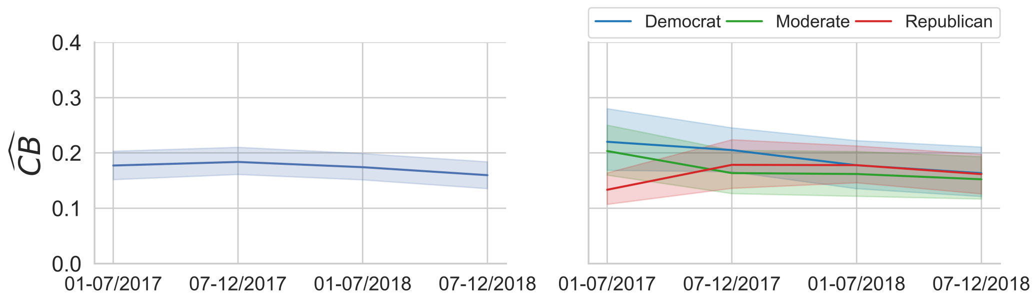


Figure 12: Confidence bound $\widehat{CB}$ mean trend of the Minority graph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

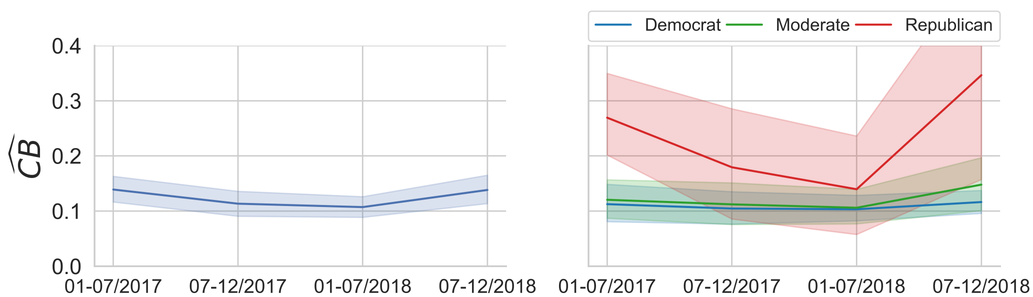


Figure 13: Confidence bound $\widehat{CB}$ mean trend of the Politics graph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

Another aspect to take into consideration is the error range of the estimated opinion. The error is computed as the absolute value of the difference between the node u observed opinion $x_u(t+1)$ at the consecutive interval $t+1$ and the estimated node opinion (Alg.1 line). The average error value for the Gun Control dataset is 0.061, for the Minority dataset is 0.0541 and for Politics 0.040. These low values suggest that the procedure implemented performs well in the estimation of the agents' opinions.

5.2 Higher order interactions

This section follows the same structure as the previous one on ordinary graphs. The Alg. 2 is applied to each dataset structured in a hypergraph configuration to compare the open-mindedness level of the three political leanings. The last section is dedicated to the influence of the discussion topics on the users' behavior.

5.2.1 Gun control

As for the pairwise graph, the first explored results are about the Gun Control dataset. For this structure, we study some additional information. The node degree represents the number of hyperlinks in which users are involved, serving as a measure of their level of activity. Higher node degrees indicate more active users, while lower degrees suggest less participation in discussions. From Figure 14 it can be seen that the Republicans tend to be in general less active, the majority of them are part of less than 50 hyperlinks. The Moderates are the most active group, they participate in more discussions with respect to the others. The Democrats have an intermediate trend between the others, in the second interval they interact less like the Republicans, otherwise, their level of activity resembles the Moderate group even though they are involved in fewer discussions.

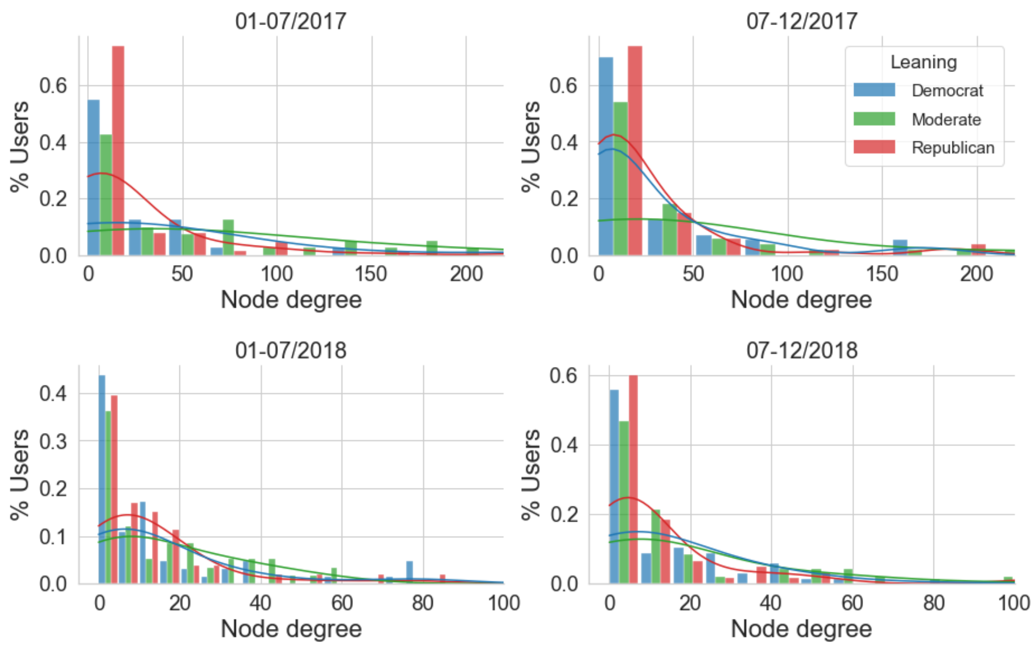


Figure 14: Distribution of the node degree for the Gun Control high-order interaction network divided by political leaning: Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Figure 15 shows the standard deviation distribution of the hyperedges for each political group. Low levels of standard deviation in the distributions suggest that users tend to interact with others who share similar opinions, leading to more homogeneous groups in terms of the opinions of the nodes belonging to the hyperedge. This finding suggests that the context of interactions is not very heterogeneous, with users participating in discussions where there is a low level of opinion diversity.

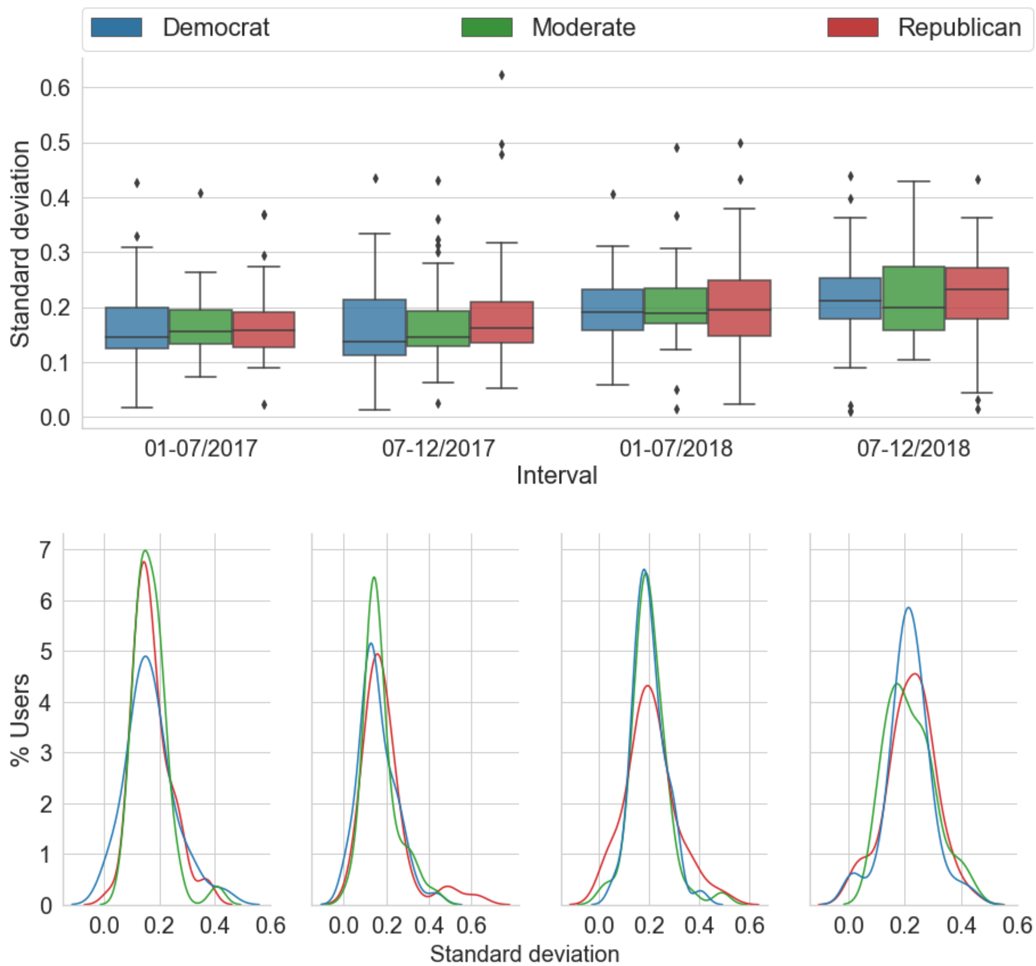
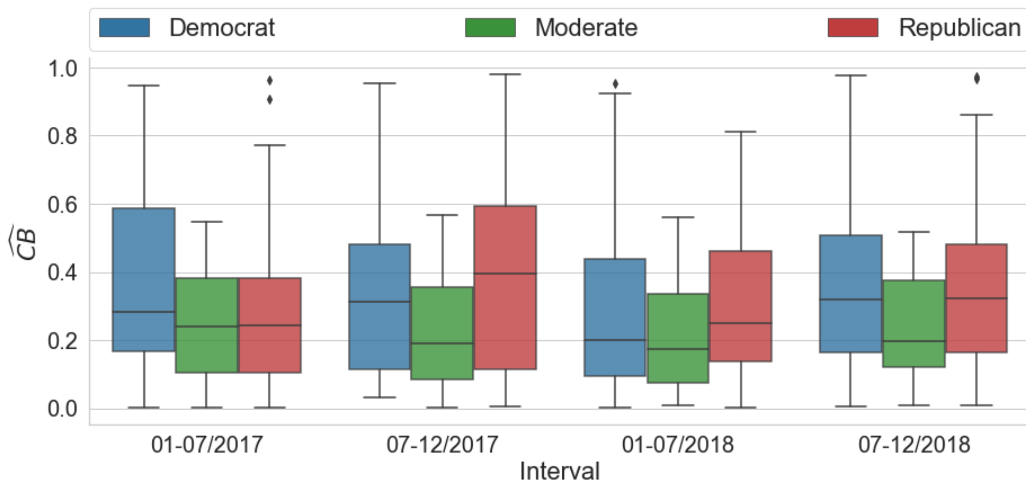
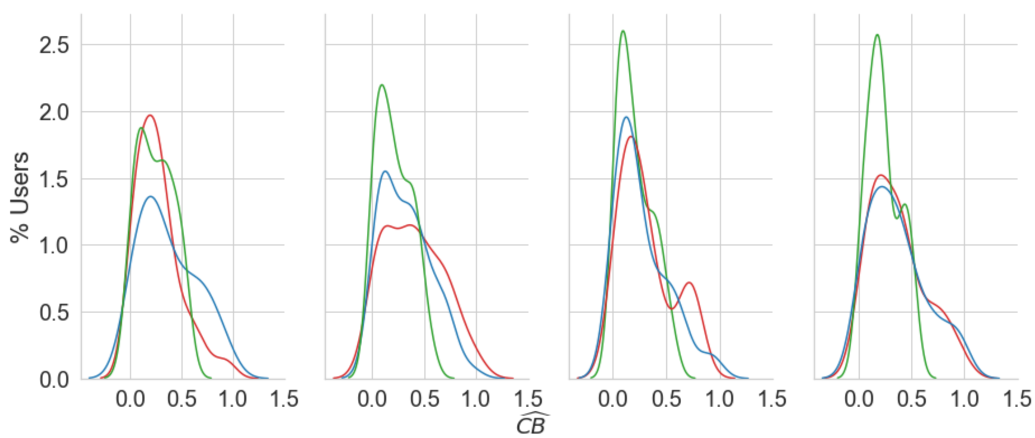


Figure 15: **Gun Control standard deviation distributions on hypergraph.** Boxplots (a) and KDE (b) of the standard deviation distributions in the Gun Control high-order interactions network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Each subpopulation (Democrat, Moderate, Republican) displays more or less the same median standard deviation. Over the considered biennium the median standard deviation increases in any case and distributions become less positively skewed, showing a peak more towards the center of the distribution. The confidence bound $\widehat{CB}$ distributions are shown in figure 16 for each political leaning during the four intervals. From the figure 15 we can see that the standard deviation distributions are less skewed w.r.t. the confidence bound $\widehat{CB}$ (figure 16), more centered and narrowed around the median. The figure 16 shows that all subpopulations have a median $\widehat{CB}$ below 0.2, which suggests that even when considering higher-order interactions, these populations can be considered "close-minded". Both population and subpopulation-level distributions are positively skewed. More in detail, Republicans are the most open-minded population, with median $\widehat{CB}$ ranging from 0.2 to 0.4 and a range of 1, covering all possible values of $\widehat{CB}$. Moderates are the most close-minded, with median $\widehat{CB}$ around 0.2 and a narrower range of values. The Moderate group displays a consistent distribution trend throughout all periods, indicating that their level of influence from online discussions on the topic remains unchanged. This group primarily consists of close-minded individuals, as previously determined by network analysis (Figure 5). The median value of their distribution is generally low, around 0.2, and over time, their confidence bound $\widehat{CB}$ level decreases further. While the two peaks observed in the pairwise network are also present in this configuration, the peak at higher values is less pronounced due to the overall lower level of open-mindedness. In contrast, the Democrat group exhibits a broader range of values for the confidence bound $\widehat{CB}$ with right-skewed distributions and higher data dispersion compared to the Moderates. Notably, during the two middle periods when four mass shootings occurred, they become more close-minded.



(a)



(b)

Figure 16: **Gun control $\widehat{CB}$ distributions estimated on hypergraphs.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Gun control high-order interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

The Republicans experience the most significant change during this period, with their distributions reaching a median value of 0.4. Additionally, the upper bound of their interquartile range shifts from 0.4 to 0.6 in the second period, resulting in an almost bell-shaped distribution. In the remaining periods, the majority of Republicans become more close-minded with right-skewed distributions.

Like in the pairwise network section, the Kolmogorov-Smirnov test is computed to have a comparison between the open-mindedness distributions of the three political leanings (table 7).

		01-07 2017	07-12 2017	01-07 2018	01-12 2018
KS-stat	Dem vs. Rep	0.231	0.162	0.145	0.078
	Dem vs. Mod	0.3	0.246	0.167	0.261
	Rep vs. Mod	0.129	0.337	0.208	0.252
P-value	Dem vs. Rep	0.12	0.411	0.513	0.977
	Dem vs. Mod	0.054	0.066	0.316	0.036
	Rep vs. Mod	0.752	0.004	0.153	0.056

Table 7: **Kolmogorov-Smirnov test** on Gun Control hypergraph network to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

The statistical test results show that in the Democrats and the Republicans comparison the null hypothesis can never be rejected (p-value > 0.05) during all the intervals and they also have the smallest distances between their distributions, especially in the last two periods. The Moderate confidence bound $\widehat{CB}$ distributions are the most statistically different when compared with the other two groups. Despite this, they are considered statistically different only in the last interval when compared to Democrats and in the second when compared to Republicans (p-value ≤ 0.05).

To see if the level of open-mindedness is strongly influenced by the node interactions context, the table 8 shows the Pearson correlation between the confidence bound $\widehat{CB}$ and the standard deviation. The table values don't show a significant correlation between the confidence bound and the deviation standard for the Democrats. For the Republicans, the highest positive correlation is found in the second interval where the group increases their open-mindedness level by interacting in contexts with heterogeneous opinions. This change could be influenced by events with a great impact on public opinion that happened during this period, for example, the Las Vegas mass shooting [3]. The same goes for the Moderates which have the highest positive correlation in the third interval.

	Pearson correlation between confidence bound $\widehat{CB}$ and standard deviation		
Interval	**Republicans**	**Moderates**	**Democrats**
01-07/2017	0.2111	0.14	-0.1218
07-12/2017	0.0328	0.3901	-0.0481
01-07/2018	0.4426	0.1428	0.2366
07-12/2018	0.2363	0.0814	0.1412

Table 8: Pearson correlation between the confidence bound $\widehat{CB}$ and the standard deviation for each political leaning in the Gun Control dataset. The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

[3] https://en.wikipedia.org/wiki/2017_Las_Vegas_shooting, last visited March 2023

5.2.2 Minorities

For the Minorities dataset, the level of activity of the three political groups is different with respect to the Gun control topic. As can be seen in the figure 17, in the first two intervals Republicans are the most active subpopulation, by participating in a higher number of discussions; in the last two, the Moderates become the more involved group. On the other hand, the majority of the Democrats are less involved in discussions about this topic.

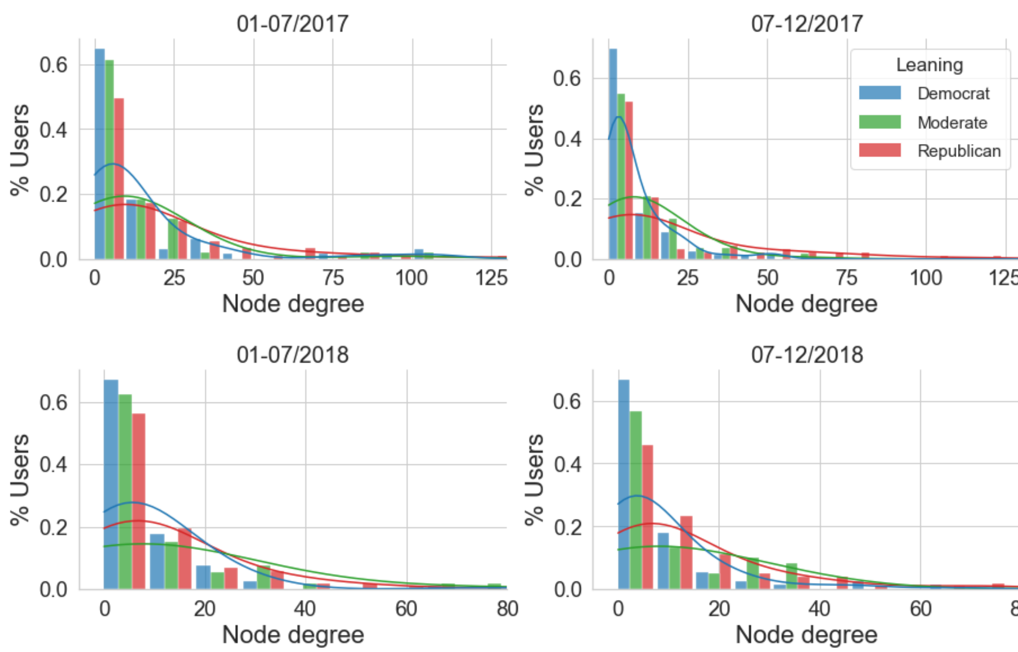


Figure 17: Distribution of the node degree for the Minority high-order interaction network divided by political leaning: Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Regarding the standard deviation analysis, even in this dataset the distributions are less skewed, but wider w.r.t. the previous dataset (figure 18).

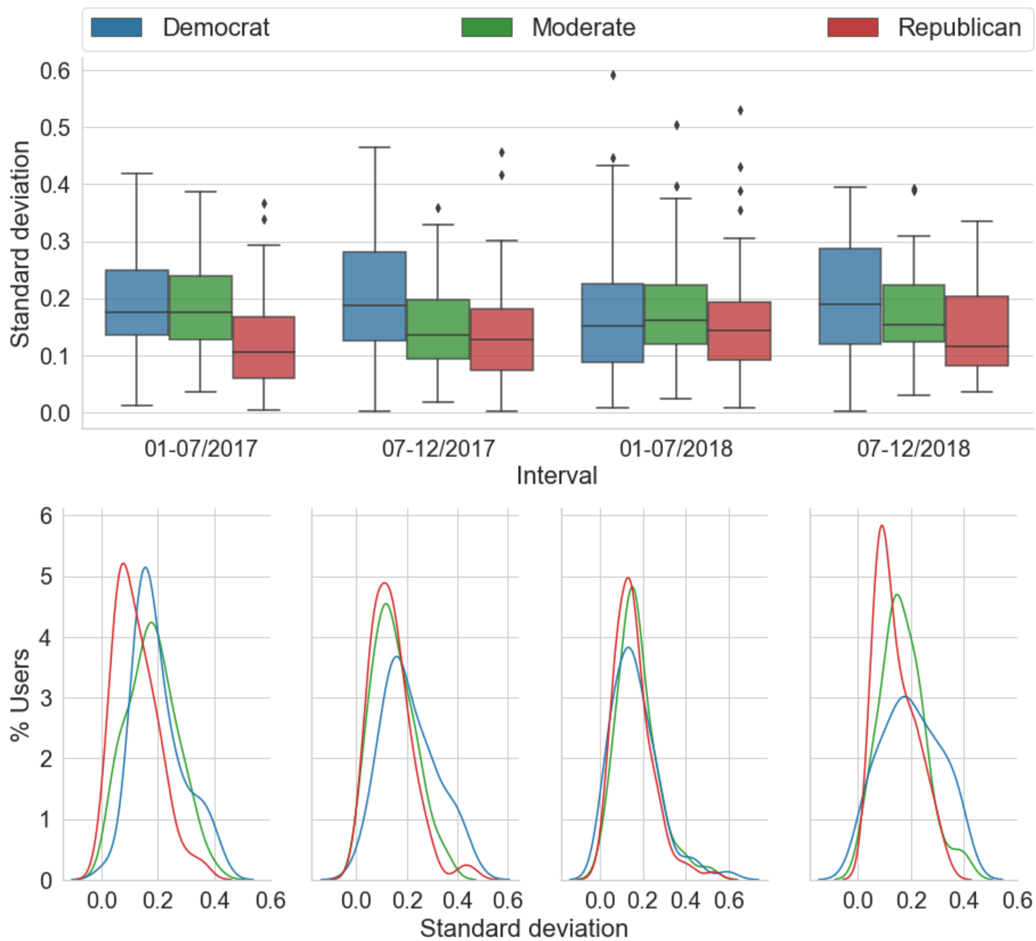
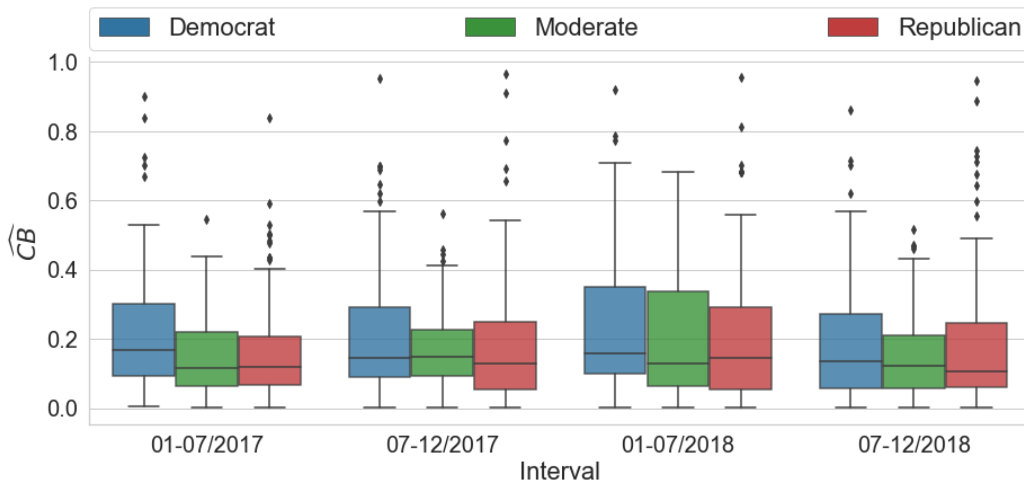


Figure 18: **Minority standard deviation distributions on hypergraph.** Boxplots (a) and KDE (b) of the standard deviation distributions in the Minority high-order interactions network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

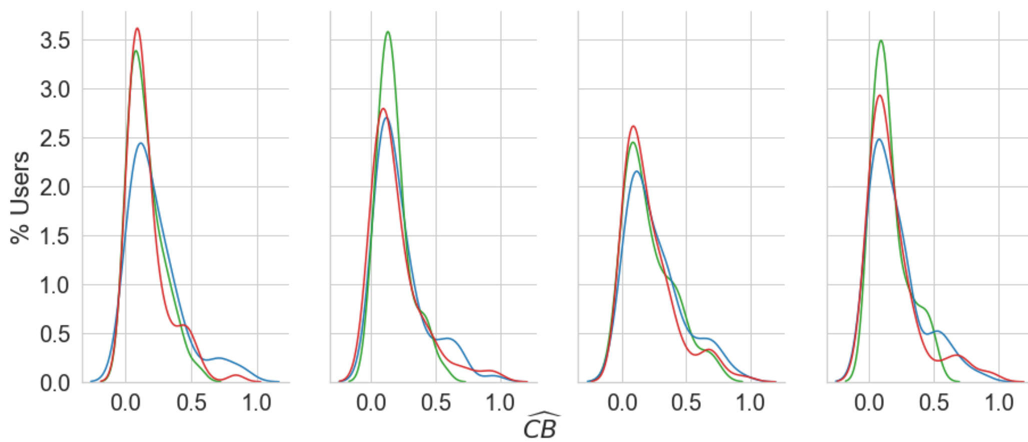
The standard deviation distribution of the Democrats changes over time, with an increase in their upper bound. As for the confidence bound $\widehat{CB}$ (figure 19), Democrats have higher standard deviation values which means that they are generally more open engaging in heterogeneous confrontations w.r.t. the other two groups.

On the opposite side, Republicans have the lowest level of standard deviation overall and they usually interact in "agreeing" contexts. Their median further decreases in the last interval. Even though Moderates interact in more heterogeneous contexts w.r.t. the Republicans, their distribution becomes narrower after the first interval, remaining in the middle between the other two political groups.

In the Minority dataset, the open-mindedness distributions of the three political leanings have similar medians in all intervals (see fig. 19). All subpopulations have right-skewed distributions with 75% of the users in each group having open-mindedness below 0.2. Democrats have the highest upper bound in all intervals while Moderates have a more centered and narrow distribution. In the third interval, all distributions are lower and wider. Republicans are the most close-minded users in this dataset. In the last interval, Democrats and Republicans distributions are similar due to an increase in the confidence bound.



(a)



(b)

Figure 19: **Minority $\widehat{CB}$ distributions estimated on hypergraphs.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Minority high-order interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

To check if the distributions are statistically different, in table 9 we report the Kolmogorov-Smirnov test results: we cannot reject the hypothesis that the three leaning empirical distributions are drawn from the same one (p-value > 0.05 in all tests).

		01-07 2017	07-12 2017	01-07 2018	07-12 2018
K-stat	Dem vs. Rep	0.208	0.147	0.183	0.119
	Dem vs. Mod	0.177	0.134	0.206	0.136
	Rep vs. Mod	0.115	0.148	0.09	0.149
P-value	Dem vs. Rep	0.061	0.287	0.086	0.551
	Dem vs. Mod	0.294	0.563	0.111	0.533
	Rep vs. Mod	0.738	0.415	0.904	0.338

Table 9: **Kolmogorov-Smirnov test** on Minority hypergraph network to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

Table 10 shows that the Pearson correlation between the Democrats confidence bound and their average hyperedge standard deviation is positive, but not high (maximum value is 0.25).

	Pearson correlation between confidence bound $\widehat{CB}$ and standard deviation		
Interval	**Republicans**	**Moderates**	**Democrats**
01-07/2017	0.0604	0.0297	0.2502
07-12/2017	0.3563	0.0913	0.1171
01-07/2018	0.3069	0.3893	0.1447
07-12/2018	0.3887	0.3352	0.1254

Table 10: Pearson correlation between the confidence bound $\widehat{CB}$ and the standard deviation for each political leaning in the Minority dataset. The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

Higher positive correlations are present between the considered measures in the last three periods for the Republican subpopulation (maximum value is 0.3887). For the Moderate group, the Pearson correlation shows a positive correlation but with higher values only in the last two intervals (maximum value is 0.389).

5.2.3 Politics

The last section focuses on the Politics dataset. This dataset is characterized by an imbalanced distribution of the agents, with only a few of them belonging to the Republican subpopulation.

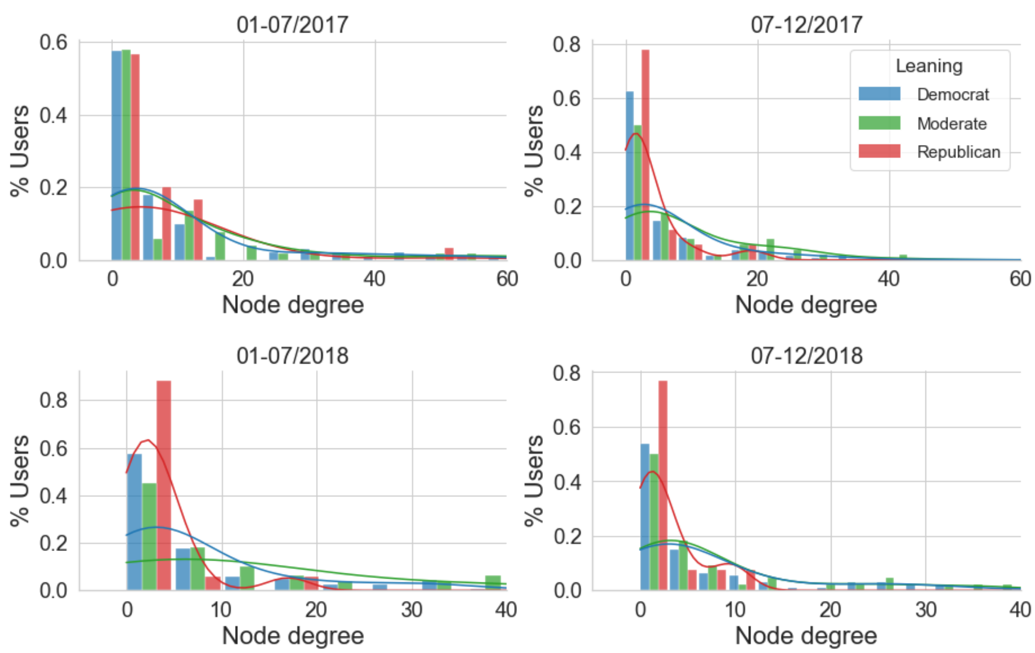
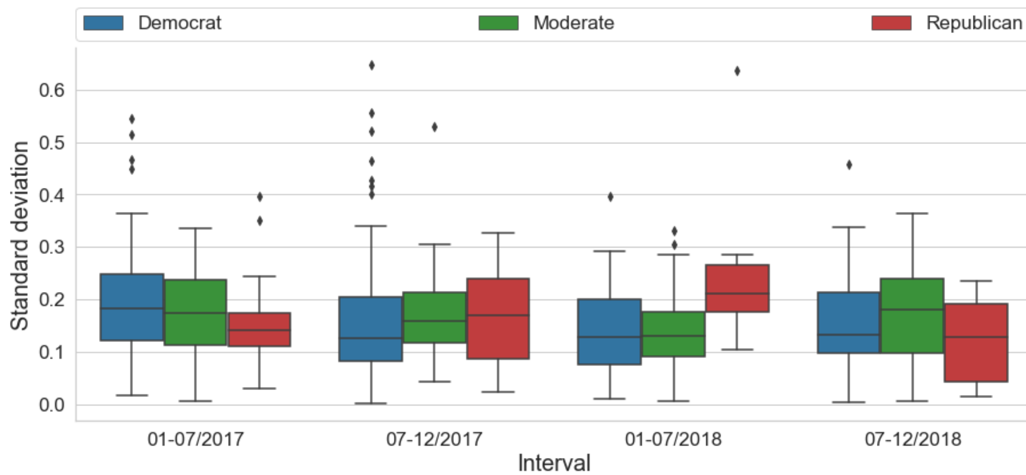


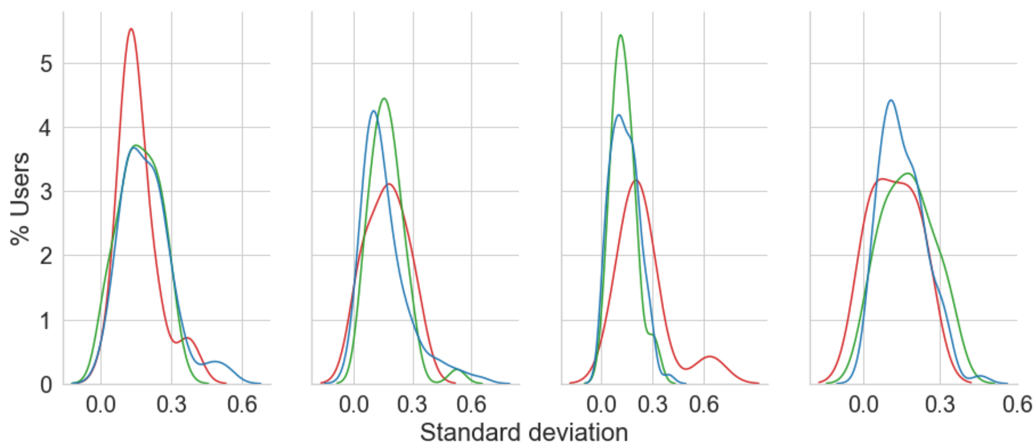
Figure 20: Distribution of the node degree for the Politics high-order interaction network divided by political leaning: Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

From figure 20 it can be seen that the users in this discussion participate in fewer interactions (lower degree), especially in the last three intervals. With respect to the previous dataset, all the nodes are less active by participating in fewer interactions.

By shifting the focus on the standard deviation analysis, even in this dataset the distributions are more centered and narrower (figure 21). In this case, the estimated confidence bound $\widehat{CB}$ of Republicans is very different w.r.t. the other groups. Moderates and the Democrats show similar trends with right-skewed distributions where the median values are always lower than 0.2. Only during the second interval, when the mid-term elections took place, the Democrats have a distribution with a longer tail and a wider interquartile range.

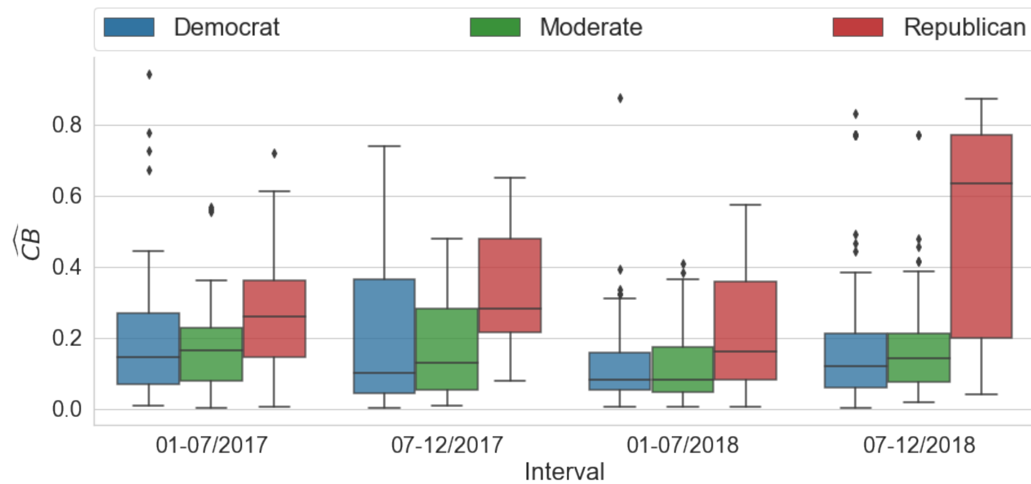


(a)

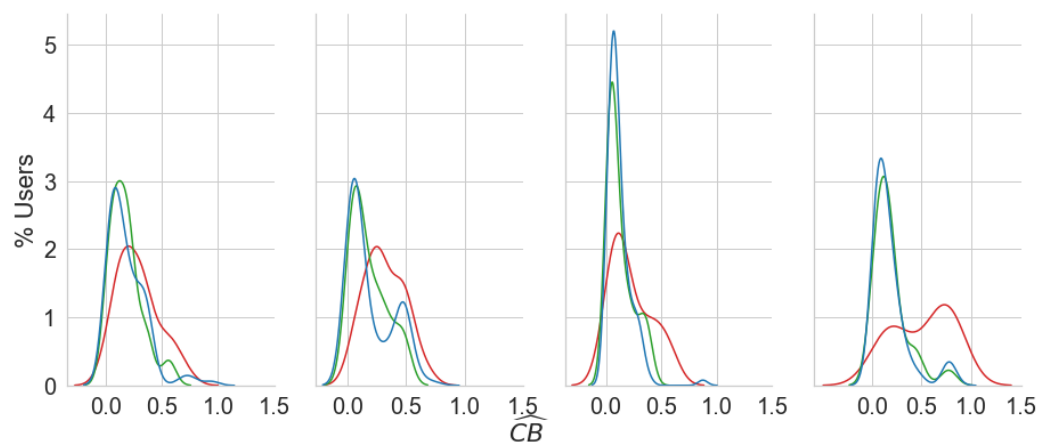


(b)

Figure 21: **Politics standard deviation distributions on hypergraph.** Boxplots (a) and KDE (b) of the standard deviation distributions in the Politics high-order interactions network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.



(a)



(b)

Figure 22: **Politics $\widehat{CB}$ distributions estimated on hypergraphs.** Boxplots (a) and KDE (b) of $\widehat{CB}$ distributions in the Politics high-order interaction network for Democrats (blue), Moderates (green) and Republicans (red). The data is aggregated by semester from January 1st, 2017 to December 31st, 2018.

The Kolmogorov-Smirnov statistical test results show (see table 11) that the comparison between Republicans and Moderates is the only one that describes the two distributions as statistically different during all the intervals. The same goes for the comparison between the Democrats and the Republicans except for the third interval (Jun-July 2018). On the other hand, the null hypothesis is never rejected in the comparison between Moderates and Democrats: the test suggests that these samples cannot be considered statistically different. The unusually broad distribution of the Republican confidence bound could be influenced by the imbalance of the political leaning distribution of the dataset. In fact figure 22 shows unusual higher values of confidence bound $\widehat{CB}$ for the Republicans w.r.t. the other users, especially in the last interval.

		01-07 2017	07-12 2017	01-07 2018	07-12 2018
K-stat	Dem vs. Rep	0.328	0.466	0.314	0.558
	Dem vs. Mod	0.114	0.128	0.148	0.149
	Rep vs. Mod	0.374	0.412	0.393	0.547
P-value	Dem vs. Rep	0.01	0.001	0.085	0.001
	Dem vs. Mod	0.705	0.522	0.384	0.416
	Rep vs. Mod	0.007	0.015	0.029	0.002

Table 11: **Kolmogorov-Smirnov test** on Politics hypergraph network to compare the $\widehat{CB}$ distributions of the three political leaning subpopulations (Republican, Moderate and Democrat)

The trend of Republicans seems different w.r.t. the others when comparing the standard deviation and the confidence bound trends. In the last interval, where they are more open-minded than before, their standard deviation is lower, unlike the others. This different trend is also highlighted in table 12 which collects the Pearson correlation between the confidence bound and the standard deviation. The Republicans have a positive correlation between the two measures in the fourth interval and a significant negative correlation in the last period. The other political groups have always positive correlations, but not very significant.

	Pearson correlation between confidence bound $\widehat{CB}$ and standard deviation		
Interval	**Republicans**	**Moderates**	**Democrats**
01-07/2017	0.3593	0.1198	0.1119
07-12/2017	0.3374	0.2639	0.1478
01-07/2018	0.5493	0.1007	0.053
07-12/2018	-0.7266	0.1006	0.2463

Table 12: Pearson correlation between the confidence bound and the standard deviation for each political leaning in the Politics dataset

5.2.4 Comparison across the three datasets

This section aims to understand how different political debate topics and the interactions of subjects with different political leanings can influence the estimated open-mindedness and its evolution over time.

Figure 25 shows the level of the aggregate estimated confidence bound $\widehat{CB}$ over the whole population (left) and divided between Republicans, Democrats and Moderates (right). As for the pairwise configuration, the highest level of aggregated open-mindedness is in the Gun control discussion: it increases in the second interval, decreases in the third, coming back at the same level as the beginning in the end. The main contribution to the increase of the aggregated distribution is given by the Republicans, which become much more open-minded as opposed to the other two groups that instead become more close-minded until the third interval. In fact, in the last period, all the users increase their confidence bound level $\widehat{CB}$. The trend of the third interval is opposite with respect to the Minorities dataset where there is a peak in this timestamp followed by a decrease. In the Politics topic, the confidence bound increases only in the last interval. In all three cases, the final aggregated level of open-mindedness is similar to the initial one. The group with the lowest level of confidence bound is the Moderates when discussing about Gun control and Minorities. In the Politics discussions, they have the same trend and values as the Democrats, while the Republicans are more open-minded on this topic. When discussing about the minority rights topic the Democrats are the ones with the highest confidence bound $\widehat{CB}$. In this case, all users from the three leanings present an increase in the third interval followed by a decrease (figure 24).

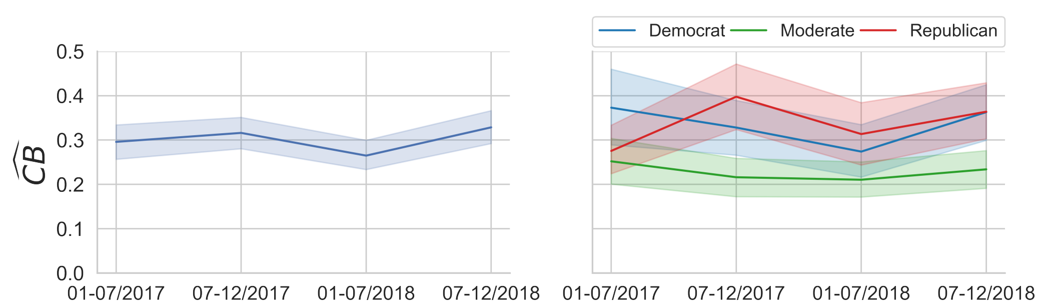


Figure 23: Confidence bound $\widehat{CB}$ mean trend of the Gun Control hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

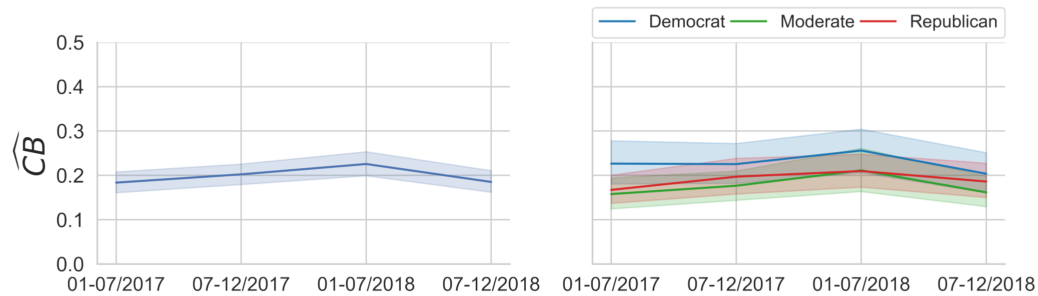


Figure 24: Confidence bound $\widehat{CB}$ mean trend of the Minority hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

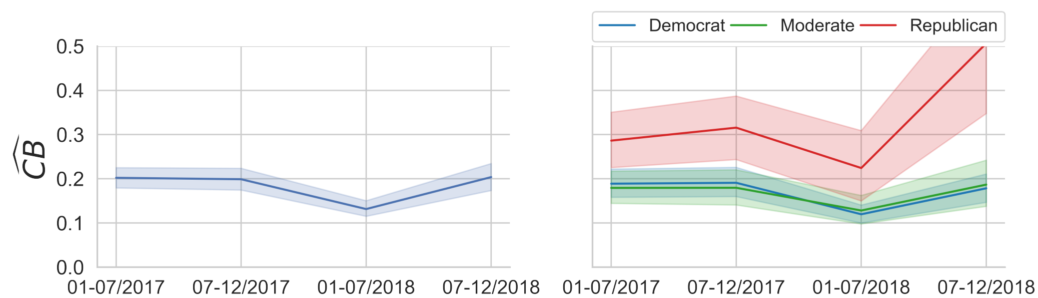


Figure 25: Confidence bound $\widehat{CB}$ mean trend of the Politics hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

The analysis continues with the understanding of the contexts where the subjects choose to interact. As can be seen in the following plots, the standard deviation trends are different in all the topics. For the Gun control discussion, the level of the standard deviation follows an increasing trend. On the other side, a more linear trend is followed for the Minorities topic, where the aggregated standard deviation value is more or less the same during all the time frames. Finally, for the Politics theme, the aggregated trend is mainly decreasing. Of the three topics, gun control is the one that reaches the highest aggregated level of the confidence bound $\widehat{CB}$ including discussions between users with different political leanings. The other two have instead the same aggregated level of standard deviation. In the Minority dataset, the highest standard deviation level is of the Democrats which is also the most open-minded group, meaning that when talking about minority rights they participate in contexts with heterogeneous opinions and are more open being influenced by discussion group with more distant average opinions. The Politics dataset is heavily influenced by the unbalanced distribution of the users political leaning the trend of the Democrat and the Moderate are very similar, like for the confidence bound 28. The Republicans show a higher level of standard deviation and open-mindedness, but this information is less trustworthy w.r.t. the other dataset, in fact, this group reaches the highest error value from the opinion estimation.

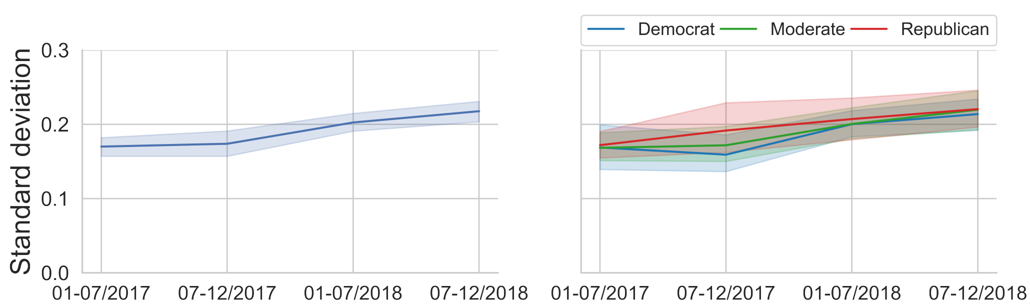


Figure 26: Standard deviation mean trend of the Gun Control hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

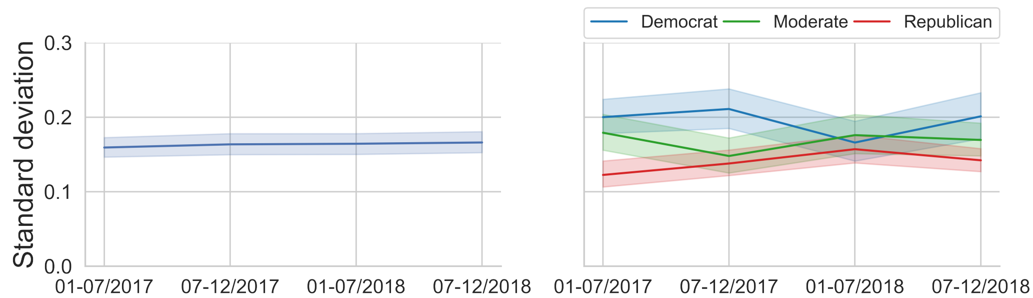


Figure 27: Standard deviation mean trend of the Minority hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

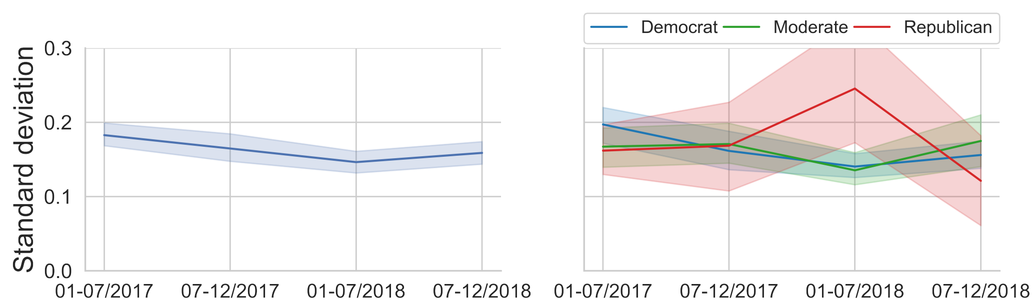


Figure 28: Standard deviation mean trend of the Politics hypergraph at an aggregated level (left plot) and for each political leanings subpopulation (right plot) from January 2017 to December 2018

Regarding the error of the estimated agents' opinions, we have that the average values are: 0.114 for the Gun Control dataset, 0.065 for the Minority dataset and 0.086 for the Politics one. Even if these values are a little higher w.r.t. the ordinary graph configuration, they are still low enough to consider the procedure well-performing.

6 Discussion

In this Chapter we discuss the similarities and the differences between the results of the procedures implemented to compute the confidence bound $\widehat{CB}$ on graphs and hypergraphs, to account for different insights that may emerge when modeling discussions as higher-order structures instead of networks.

6.1 Comparison between different underlying structures

Starting from Gun control, the results from the two configurations are quite different comparing the figures 11 and 23. In both cases, the Moderates are the most close-minded subpopulation, but in the hypergraph, their confidence bound $\widehat{CB}$ is more limited, concentrated around a mean value of 0.2 and presents a shorter interquartile range, w.r.t. the ordinary graph. The opposite results are found for the Democrats, for which the interquartile range is shifted on higher confidence bound $\widehat{CB}$ values in the hypergraph configuration. In fact, comparing figure 11 and 23, at the aggregated level the procedure on the ordinary graph shows general lower level of confidence bound. In the pairwise graph, the lowest confidence bound $\widehat{CB}$ values, are always of the Democrats or the Republicans while on the hypergraph this position is occupied by the Moderate. In fact, from the figure 5, it can be seen that for the pairwise interactions the confidence bound $\widehat{CB}$ trends of the three political groups are more similar while for the hypergraph, the Moderates have a very different behavior w.r.t. the other groups with a lower level of open-mindedness. Even if at different levels, the trends of the distributions are more or less the same for the Republicans and the Moderates. This means that for these two groups, both procedures capture the same behavior.

Regarding the estimation error, results on the networks show lower error with respect to the hypergraph configuration.

For the Minority dataset, looking at figures 12 and 24 we can see that only Republicans have similar behaviors. The focus on the political leanings displays that in both procedures the Democrats are the less close-minded, followed by the Republicans and then the Moderates. As for the Gun control dataset, the aggregated trends show higher confidence bound $\widehat{CB}$ level for the hypergraph configuration and the same goes for the error.

The last dataset, the Politics, matches the same considerations already expressed for the other datasets, the higher aggregated confidence bound $\widehat{CB}$ are computed from the procedure applied on the hypergraphs which has also a higher level of estimated opinion error. In this case, the trends of the three political leanings found in both configurations are very similar (figure 13 and figure 25). Overall it can be seen that the hypergraph configuration tends to produce higher values of confidence bound $\widehat{CB}$ with respect to networks, with which it is instead possible to reach the lower level of estimation error. This is due to the fact that the context of group interactions modeled by the hyperedges amplifies the influence of nodes with more distant opinions during the opinion estimation process.

7 Conclusions

This thesis adopts a data-driven approach to estimate the level of individual open-mindedness in different controversial debates on Reddit about U.S. politics. The data collected covers three main discussion topics: gun control legislation, minority discrimination, and general socio-politics arguments [4]. The data time window goes from January 2017 to December 2018, regarding the first two years of Trump's presidency. The data collected are organized in two different network structures: graphs (pairwise interactions) and hypergraphs (high-order interactions). A real-valued opinion is assigned to each node of the network, representing their political leaning, identifying users as Republicans, Democrats or Moderates.

Using as a reference the Deffuant-Weisbuch model [9], which is one of the bounded confidence opinion dynamics models, two different methodologies are implemented for the two modeling configurations. Both procedures aim to estimate one of the fundamental parameters of the Deffuant-Weisbuch model, i.e, the confidence bound $\widehat{CB}$, which represents the maximum opinion distance between users' opinions for active interaction between them. Since this parameter is usually assumed when testing simulations on graph models, this thesis aims to define a framework to estimate it from real interaction data.

The results on networks show that the majority of Reddit users are closed minded having a low level of confidence bound, which prevents them to reach consensus and causing opinion fragmentation [9]. An interesting factor is that between the political leaning groups, usually the most closed-minded users belong to the Moderate subpopulation rather than the other two which have more extreme political ideologies. Moreover, the discussion topic has a great influence on the way the agents interact, for example when debating about gun control, the Republicans have in general higher level of confidence bound $\widehat{CB}$, while for minority rights, the Democrats take that position. Focusing on different discussion topics allows also to see how historical events influence the confidence bound level $\widehat{CB}$, especially in controversial topics like the gun control in the U.S. The different open-minded distributions for different topics show the multiple facets which characterize the agents' opinions and their interaction choices.

Results from hypergraphs show that the confidence bound $\widehat{CB}$ of the political leaning groups is generally low, preventing populations to reach consensus. Observations here are the same drown in the previous setting; even though the confidence bound $\widehat{CB}$ levels are a little higher in this configuration, they aren't enough to overcome opinion fragmentation in the network, according to the theoretical framework.

Overall, the estimation error representing the difference between the estimated opinion and the observed one, shows that both methodologies are good approximations of the interaction influence in the networks.

7.1 Limitations

The main limitations of this thesis are linked to the analysis approach adopted presented in section 2.2.3. On one hand, the data-driven approach is very useful to understand the pattern of the data, to customize the possible procedure applications, and also, like in this case, to study if and how similar methodologies applied on different network structures can bring similar or different conclusions. On the other hand, the results are strictly linked with the context of the analysis and can't be generalized or applied in other scenarios like in wider or dissimilar social, cultural or historical backgrounds. For example, the open-mindedness trends defined in this work are specific for American users and can't be generalized to European ones due to the complexity and peculiarity of the social settings. Moreover, as can be seen in Chapter 5 the results show that the procedures give different insights also based on the topic of discussions, which can be less controversial in other cultures or social contexts.

The complexity is not just linked to the analysis approach adopted but also to the Opinion Dynamic field itself. The majority of the work in this area is model-driven due to the difficulty to collect data, but also to estimate with a numeric value the users' opinions. The process of opinion estimation necessarily implies a simplification of the various range of opinions that a subject can hold, by reducing them to numeric variables implies inevitably a limit in the representation of the reality.

Linked to the data availability, the other limitation is that it could be more useful if the collected information covers longer periods of time to better understand if, in the long run, the two methodologies implemented show similar results or opposite insights also w.r.t. important historical and political events like presidential elections, new legislation and other events.

7.2 Future works

To overcome the limitations presented in the previous section, it can be useful to continue to collect and process new data from the same data sources and with the same processes to be able to have an insight in the long run. Having a longer time frame can be helpful not only to understand the network dynamic evolution but also in which ways new events influence the results and if in the long run, the two procedures can detect the same changes and define the same evolution patterns.

Moreover, it could be interesting to define a network structure (for example Barabasi-Albert model [21]) with the same characteristics of the real network of this thesis, to test simulations using different level of confidence bound to understand the possible future polarization or consensus of opinions.

References

[1] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *Proceedings of the international aaai conference on web and social media*, vol. 5, no. 1, 2011, pp. 89–96.

[2] A. J. Morales, J. Borondo, J. C. Losada, and R. M. Benito, "Measuring political polarization: Twitter shows the two sides of venezuela," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 3, p. 033114, 2015.

[3] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[4] V. Morini, L. Pollacci, and G. Rossetti, "Capturing political polarization of reddit submissions in the trump era." in *SEBD*, 2020, pp. 80–87.

[5] E. Pariser, *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.

[6] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[7] Z. Kunda, "The case for motivated reasoning." *Psychological bulletin*, vol. 108, no. 3, p. 480, 1990.

[8] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[9] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 3, no. 01n04, pp. 87–98, 2000.

[10] S. Verducci, "Critical thinking and open-mindedness in polarized times," *Encounters in Theory and History of Education*, vol. 20, no. 1, pp. 6–23, 2019.

[11] C. R. Sunstein, *Republic.Com 2.0*. USA: Princeton University Press, 2007.

[12] A. Sîrbu, V. Loreto, V. D. Servedio, and F. Tria, "Opinion dynamics: models, extensions and external effects," *Participatory sensing, opinions and collective awareness*, pp. 363–401, 2017.

[13] A. F. Peralta, J. Kertész, and G. Iñiguez, "Opinion dynamics in social networks: From models to data," *arXiv preprint arXiv:2201.01322*, 2022.

[14] M. Maes and L. Bischofberger, "Will the personalization of online social networks foster opinion polarization?" *Available at SSRN 2553436*, 2015.

[15] V. Pansanella, V. Morini, T. Squartini, and G. Rossetti, "Change my mind: Data driven estimate of open-mindedness from political discussions," in *Complex Networks and Their Applications XI: Proceedings of The Eleventh International Conference on Complex Networks and Their Applications: COMPLEX NETWORKS 2022—Volume 1*. Springer, 2023, pp. 86–97.

[16] G. Weisbuch, "Bounded confidence and social networks," *The European Physical Journal B*, vol. 38, pp. 339–343, 2004.

[17] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

[18] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[19] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J. P. Nadal, A. Sanchez *et al.*, "Manifesto of computational social science," *The European Physical Journal Special Topics*, vol. 214, pp. 325–346, 2012.

[20] L. Lovász, *Large networks and graph limits*. American Mathematical Soc., 2012, vol. 60.

[21] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[22] D. J. d. S. Price, "The scientific foundations of science policy," *Nature*, vol. 206, pp. 233–238, 1965.

[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[24] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[25] P. Holme and J. Saramäki, "Temporal networks," *Physics reports*, vol. 519, no. 3, pp. 97–125, 2012.

[26] T. Viard, M. Latapy, and C. Magnien, "Computing maximal cliques in link streams," *Theoretical Computer Science*, vol. 609, pp. 245–252, 2016.

[27] M. Latapy, T. Viard, and C. Magnien, "Stream graphs and link streams for the modeling of interactions over time," *Social Network Analysis and Mining*, vol. 8, pp. 1–29, 2018.

[28] M. Coscia, "The atlas for the aspiring network scientist," *arXiv preprint arXiv:2101.00863*, 2021.

[29] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, and G. Petri, "Networks beyond pairwise interactions: structure and dynamics," *Physics Reports*, vol. 874, pp. 1–92, 2020.

[30] P. Clifford and A. Sudbury, “A model for spatial conflict,” *Biometrika*, vol. 60, no. 3, pp. 581–588, 1973.

[31] R. A. Holley and T. M. Liggett, “Ergodic theorems for weakly interacting infinite systems and the voter model,” *The annals of probability*, pp. 643–663, 1975.

[32] S. Galam, “Minority opinion spreading in random geometry,” *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 25, pp. 403–406, 2002.

[33] K. Sznajd-Weron and J. Sznajd, “Opinion evolution in closed community,” *International Journal of Modern Physics C*, vol. 11, no. 06, pp. 1157–1165, 2000.

[34] R. Hegselmann and U. Krause, “Opinion dynamics driven by various ways of averaging,” *Computational Economics*, vol. 25, pp. 381–405, 2005.

[35] A. Sîrbu, D. Pedreschi, F. Giannotti, and J. Kertész, “Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model,” *PloS one*, vol. 14, no. 3, p. e0213246, 2019.

[36] H. Schawe and L. Hernández, “Higher order interactions destroy phase transitions in deffuant opinion dynamics model,” *Communications Physics*, vol. 5, no. 1, p. 32, 2022.

[37] H. Rainer and U. Krause, "Opinion dynamics and bounded confidence: Models, analysis and simulation," *Journal of Artificial Societies and Social Simulation*, vol. 5, no. 3, 2002.

[38] A. Hickok, Y. Kureh, H. Z. Brooks, M. Feng, and M. A. Porter, "A bounded-confidence model of opinion dynamics on hypergraphs," *SIAM Journal on Applied Dynamical Systems*, vol. 21, no. 1, pp. 1–32, 2022.

[39] A. Vendeville, B. Guedj, and S. Zhou, "Forecasting elections results via the voter model with stubborn nodes," *Applied Network Science*, vol. 6, 2021. [Online]. Available: https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00342-7

[40] L. Li, A. Scaglione, A. Swami, and Q. Zhao, "Consensus, polarization and clustering of opinions in social networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, pp. 1072–1083, 2013.

[41] P. Sobkowicz, "Quantitative agent based model of opinion dynamics: Polish elections of 2015," *PLoS ONE*, vol. 11, 2016.

[42] A. Kononovicius, "Empirical analysis and agent-based modeling of the lithuanian parliamentary elections," *Complex.*, vol. 2017, pp. 7 354 642:1–7 354 642:15, 2017.

[43] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin, "Information gerrymandering and undemocratic decisions," *Nature*, vol. 573, no. 7772, pp. 117–121, 2019.

[44] F. Xiong and Y. Liu, "Opinion formation on social media: an empirical approach," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 1, p. 013130, 2014.

[45] C. Monti, G. De Francisci Morales, and F. Bonchi, "Learning opinion dynamics from social traces," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 764–773.

[46] W. Jager and F. Amblard, "Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change," *Computational & Mathematical Organization Theory*, vol. 10, pp. 295–303, 2005.

[47] O. A. Sichani and M. Jalili, "Inference of hidden social power through opinion formation in complex networks," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 3, pp. 154–164, 2017.

[48] A. De, I. Valera, N. Ganguly, S. Bhattacharya, and M. G. Rodriguez, "Learning and forecasting opinion dynamics in social networks," in *Advances in neural information processing systems*, 2016, pp. 397–405.

[49] B. Kulkarni, S. Agarwal, A. De, S. Bhattacharya, and N. Ganguly, "Slant+: A nonlinear model for opinion dynamics in social networks," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 931–936.

[50] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 830–839.

Ringraziamenti

Vorrei riservare quest'ultima pagina per i ringraziamenti finali.
Innanzitutto, desidero ringraziare i relatori, il Prof. Rossetti Giulio e la Dr. Pansanella Valentina, per avermi dato l'opportunità di svolgere questa tesi, per la loro costante disponibilità e competenza, essenziali per la riuscita del presente lavoro.
Ringrazio i miei genitori, zii e nonni per la loro fiducia e incoraggiamento continuo. Il loro sostegno incondizionato ha avuto un ruolo essenziale nel raggiungimento di questo mio traguardo.
Un grazie speciale a mio fratello Enrico che mi ha sempre supportata alleggerendo i momenti di tensione e sconforto con la serenità che lo contraddistingue.
Infine desidero ringraziare le amiche di sempre Laura, Anna e Lisa, e gli amici conosciuti durante il percorso universitario, Alberto, Marco e Marta con i quali ho condiviso momenti di studio e divertimento.

Con sincera gratitudine,

Giulia Ferro