Causality-based Social Media Analysis for Normal Users Credibility Assessment in a Political Crisis

Ahmed Abouzeid, Ole.Christoffer Granmo, Christian Webersik, Morten Goodwin
University of Agder
Grimstad, Norway
{ahmed.abouzeid, ole.granmo, christian.webersik, morten.goodwin}@uia.no

***Abstract*—Information trustworthiness assessment on political social media discussions is crucial to maintain the order of society, especially during emergent situations. The polarity nature of political topics and the echo chamber effect by social media platforms allow for a deceptive and a dividing environment. During a political crisis, a vast amount of information is being propagated on social media, that leads up to a high level of polarization and deception by the beneficial parties. The traditional approaches to tackling misinformation on social media usually lack a comprehensive problem definition due to its complication. This paper proposes a probabilistic graphical model as a theoretical view on the problem of normal users credibility on social media during a political crisis, where polarization and deception are keys properties. Such noisy signals dramatically influence any attempts for misinformation detection. Hence, we introduce a causal Bayesian network, inspired by the potential main entities that would be part of the process dynamics. Our methodology examines the problem solution in a causal manner which considers the task of misinformation detection as a question of cause and effect rather than just a classification task. Our causality-based approach provides a practical road map for some sub-problems in real-world scenarios such as individual polarization level, misinformation detection, and sensitivity analysis of the problem. Moreover, it facilitates intervention simulations which would unveil both positive and negative effects on the deception level over the network.**

I. Introduction

Nowadays, social media are an essential part of humans everyday life. The notable increase of the number of users, the ease and cheap cost of information sharing, and the consecutive technological enhancements of different social media platforms, have indeed boosted social media to be a tough competitor to traditional news outlets. The great benefit of social media does not lie only in news circulation, but further contexts of propagating it have been introduced. These modern paradigms of fast and cross-distance social interaction have allowed additional perspectives on how people are responding to the news. For example, emotions, questions, and disapproval from eyewitnesses have became part of the process.

Emergencies are not an exception of how people would depend on social media. During critical scenarios like political rebellions, terrorist attacks, and disasters caused by natural hazards, a significant amount of information is being propagated. In such circumstances, authorities and citizens construct various usage patterns of information through social media [1]. For instance, citizens would like to get updated by following authorities verified Facebook pages or Twitter accounts. Also, authorities might rely on information disseminated by citizens to feed up their emergency management systems in order to support decisions, since people could act as eyewitnesses. Moreover, citizens would interact together to enable more information diffusion, to express emotions, or to offer help. It has been observed from previous studies that there is a major challenge in all these patterns, which is the information manipulation and the lack of trust between citizens and authorities or between citizens themselves [2], [3], [4].

Information veracity assessment on social media is a critical topic because of how misleading news could affect the social order and the recovery from an emergency. Besides, the lack of trust between social media consumers threatens social media to serve as a reliable source of information, and waste all the technological efforts that have been previously accomplished. Moreover, how people are more likely to disseminate information regardless its correctness has its roots in psychological and social literature [5], [6]. Fortunately, with the growing number of consumers and how they depend on social media, users are playing a fundamental role in questioning and verifying received information, which can be viewed as a self-defense mechanism. Despite the contradiction of how both individuals and societies could positively or negatively shape information credibility, this self-defense mechanism of social media reveals the feasibility of overcoming such difficulty. Therefore, the study of information trustworthiness on social media has brought more focus in recent years.

Information accuracy dilemma on social media can be broken into multiple sub-problems: *rumor detection*, *cyborg/ trolls/ social bots detection*, and *fake news detection*. In rumor detection, a rumor could be either correct or incorrect. Commonly, a rumor is created during an emergency and due to the absence of a reinforced report from official entities. Some literature defines a rumor as a possibility to be either true or false [7]. Ref. [8] defined a rumor as an item of information which is deemed to be false. A potential cause of incorrect or inconsiderable information is social media fake accounts. Hence, trolls, cyborg, and social bots detection have been studied in some literature in the preceding years [9], [10], [11]. In the context of social media, trolls are deceptive accounts ran by a human whom purpose is to motivate the others to an emotional reaction. On the other hand, cyborgs are a semi-automated accounts which objectively try to spread fake information. Social bots are usually ran by a computer program and used in many cases like advertising and fake news circulation. Fake news detection is the process of discovering false news. Either it was misinformation or disinformation. Misinformation refers to the unintentionally spreading of false information. On the contrary, disinformation

ISSN 2305-7254

is purposely circulating of fake news and usually adopted in political propaganda or in financial manipulations [12], [13]. In the rest of this paper, we will use the term Misinformation to refer to any fake news, regardless of the intentions.

Our theoretical study focuses on the circulation of misinformation in political emergencies like revolutions and uprisings, where corrupted regimes and citizens might confuse and mislead the public by disseminating deceptive content. One of the recently revealed methods on social media misinformation is the propagation-based method which considers that more information trustworthiness evidence to be retrieved from a majority of eyewitnesses or verified accounts [14]. In propagation-based methods, credibility networks are built to employ optimization techniques over different pieces of news giving the underlying point of view. Mining different viewpoints and reactions to news is referred to as Stance Detection [15], [16]. However, in a political crisis, everyone is biased with their opinions and reactions to other opinions or shared news [17]. Hence, such context is challenging the assumption that we can unveil information credibility by investigating different opinions and find out what the majority of people are believing in. On the other hand, it would be less complicated during disasters caused by natural hazards because people are less biased and the available opinions would be easily trusted.

A. Contribution and Paper Organization

This paper focuses on the problem of normal users content credibility assessment from the perspective of cause and effect as an interpretation of some evidence during the investigation. The paper studies a potential novel approach to the problem by engaging a theoretical foundation from Bayesian analysis and causal inference [18], [19]. The study challenges the assessment of different opinions trustworthiness about a specific claim. For that, we propose a probabilistic graphical model based on a causal Bayesian network to reason about possible causes and effects within the dynamics of information propagation on social media platforms. Our proposed method tries to solve the challenge of the unreliable opinion-based solutions in polarized scenarios by calculating a posterior marginal probability of the trustworthiness degree of opinions after obtaining some evidence. Our research contributions are summarized as follows.

- explain the capabilities of both predictive and diagnostic analysis on Bayesian networks to infer about the trustworthiness of an opinion and estimating polarization level and other unknown information, given some evidence and observed causes and effects;
- illustrate how causal-based social media analysis opens the road to a potential novel approach for misinformation detection with the three layers of causal inference;

Although the different opinions and reactions to the news are taken into consideration along with the detected biased communities and deceptive accounts in a social network. However, it is important to highlight that the stance knowledge extraction task and community deceptive accounts detection techniques are not the focus of this research paper. The remainder of the paper is organized as follows. Section 2 gives a summary of the related work. Section 3 demonstrates the problem statement and notations. Our causality-based approach is explained in more details in section 4. Our proposed methodology provided by a toy example is explained in section 5. Finally, we conclude the whole paper and discuss our future work in section 6.

II. Related Work

Research efforts on social media misinformation have studied the construction of prediction models for a misinformation classification task [2]. These models can be categorized into two classes: *content-based models*, and *context-based models*. Content-based models are divided into two main approaches: *knowledge-based*, and *style-based*. Knowledge-based methods propose examining external sources to fact-check the suspected information [20]. Various approaches could be applied for fact-checking as it could be automated or managed by human experts or crowd-sourcing. Computational fact-checking methods usually use either open web or a structured knowledge graph [21]. A knowledge graph is a structured network topology which could be constructed from the open web such as DBpedia and Google Relation Extraction Corpus. A fact-checking procedure is adopting a knowledge graph in order to infer about facts on its graph to verify information by exploring evidence from the external information source [21].

Style-based methods try to capture information manipulators by their writing style. Style-based methods can be categorized into two main classes: *deception-oriented*, and *objectivity-oriented* [2]. Earlier studies from forensic psychology investigated the credibility and manipulation of statements [22]. Such studies motivated the deceptive-oriented methods to detect misinformation. Deep neural network models, such as Convolutional Neural Networks (CNN), have been applied to classify deceptive contents according to their deceptive attitudes [23]. Objectivity-oriented refers to the manipulation of news by decreasing or hiding a key piece of information. Such scenarios are likely to happen in political emergencies and political manipulation campaigns. Linguistic-based features were used to detect objectively manipulated news articles [24].

Context-based models have two main approaches: *stance-based*, and *propagation-based*. Stance-based studies users reactions on the news. Some work proposed a bipartite network of users and Facebook posts using the ”like” stance [16]. This network was used for a semi-supervised probabilistic model to predict how likely a Facebook post is a hoax. Propagation-based approach focuses on people opinions on social media, it relies on the assumption that information credibility is highly related to the sincerity of social media opinions in relevant contents. Propagation-based approach attempts to infer if there is any conflict in the shared information by exploring other circulated details associated with a particular topic. Two types of propagation networks could be built: *homogeneous* and *heterogeneous* credibility networks [2]. Homogeneous credibility networks consist of a single kind of entities, such as posts or events [15]. Heterogeneous credibility networks connect different types of entities, such as posts, and sub-events [25]. These networks performs an optimization task on their graphs to conclude the veracity of the information.

It is acknowledged that the problem of misinformation in political situations cannot be solved by only applying any state-of-the-art technology in similar domains. For instance, stance detection and text-based solutions can just act as a first phase for a complicated pipeline. That is because in polarized political scenarios, the definition of fake news is relative, due to the different perspectives each sub-group would have. Therefore, what is really misinformation differs from the perceived false content. Recent efforts in studying the relationship between fake information and political polarization have revealed a correlation between polarization and what people consider as fake news on Twitter [26]. That claims an obstacle in the combat of misinformation detection on social media since concepts like biased opinions, actual fake information, relatively fake contents can be easily confused because of such correlation.

The polarization caused by both social media platforms and human nature threatens the reliability of opinion-based misinformation detection methods. In general, many social network community detection algorithms have been adopted [27]. Previous studies tried to tackle such problem by assuming that if we enforced more information diversity to each social bubble, it would reduce the polarization since the latter is an effect of the lack of information diversity itself [28]. Other previous work aimed to detect these communities and identify them as sub-networks or similar connected nodes in the social graph by analyzing the network cohesion [29]. One of the common real-world networks in community detection is Zachary's karate club which is a real example of a social network of 34 members (nodes) in a karate club and usually used as a benchmark dataset to evaluate community detection algorithms as well [30]. One of the recent contributions was an incremental method to detect communities in dynamic evolving social networks which was motivated by how previous community detection methods were static [31].

A similar concept to misinformation on social media is disease diagnosis and detection, both issues are putting people's lives on danger and they have symptoms and causes. Both also can spread among societies and their sub-communities. One of the most advanced techniques in modern medical diagnosis is the Bayesian Network (BN) [32]. BNs are probabilistic graphical models used to represent conditional relationships between random variables (graph nodes). These random variables could represent both evidence and quires which we aim to reason and infer about. The relations in BNs can be modeled as causal relations which are more suitable for problems when the causes and effects are the core of the situation dynamics.

The problem of information veracity assessment on social media is intersected with many other tasks in the domain of social media analysis during disasters and other related contexts. Hence, it is important to highlight that in general, research community aims to extract knowledge from social media during crisis but there are differences between each sub-task. Knowledge extraction can be applied for sentiment analysis to explain the social behaviour of citizens during different stages of a crisis [33]. Opinion extraction used for news credibility tasks or political analysis [34]. Geo-location extraction tasks are being approached during disasters caused by natural hazards [33]. Hate speech towards certain groups of people which commonly increases during refugee crisis or extremely polarized political crisis [35].

III. Problem

A. Misinformation Definition

We aim to define the problem of normal users content credibility assessment on social media during a political crisis as a cause and effect problem instead of a classification task. The reason behind such definition is that an ordinary classification approach would not provide a complete control of such critical issue in our societies. On the other hand, an intervention view could unveil the root causes or suggest strategies to prevent misinformation. For that, we define misinformation propagation in terms of both predictive and diagnostic analysis tasks where causal inference approach is strongly followed. Misinformation could be viewed as a disease and the task is to understand when that disease occurs, and why it happens, and how to stop such issue in advance.

In the process of misinformation spreading, individuals approval to deceptive contents, and information shared by extremely polarized persons, could be considered symptoms of the deception phenomenon. Fig. 1 shows misinformation analysis causal-inspired solution framework. The framework declares how the stance detection, polarization measures, variety of social content a user is exposed to, deception information, and causal relations are considered as evidence to be collected in order to compute the trustworthiness of a user opinion. Moreover, an advanced causal analysis would be applied, such as intervention and sensitivity analysis to provide more confidence and insights about the inference or to hopefully suggest defensive strategies [19].

The definition of deception is critical to our proposed causal approach. We consider a political crisis as an environment where trolls, cyborgs, and deceptive social bots are trying to manipulate the public and motivate them to a specific reaction (stance). We differentiate between the collected deception information (deceptive accounts) and the unknown credibility of normal users. The latter is our focus in this study as we believe normal users are the threatening carriers of a deception disease in political discussions. Hence, a strategy to only detect trolls, cyborgs, and deceptive social bots is not sufficient in our

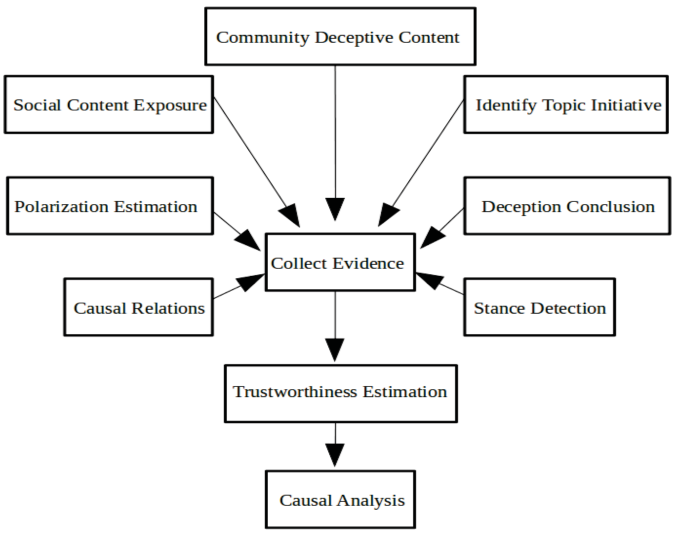

Fig. 1. Misinformation analysis causal-inspired solution framework

opinion.

As Fig. 1 indicates, we define a community deceptive content and a deception conclusion as two different things. Community deceptive content are all deceptive accounts in all biased communities over the network. For instance, right-wing trolls would manipulate the right-biased users to agree on a certain topic, on the other hand, left-wing cyborgs would defend that by propagating a refusal stance, however, both left/right-wings might share the same stance in some cases. We consider users as less trusted if they agreed on a common deceptive stance which was disseminated by all detected community deceptive content (left/right-wings). Although, in most cases, these community manipulating accounts would disagree with each others, therefore, a conclusion of the deception should be defined. Such conclusion means which opinion is considered less trustworthy and which could be dealt with as a defensive mechanism. To set a conclusion and draw the boundary lines between the differences in deceptive content stances, one more causal entity should be introduced, that is the topic initiative.

Topic initiative is defined as which biased party initially circulated the stance about the claim. For example, initially sharing something with an agreement or disagreement on it. The topic initiative would help to conclude the actual deceptive stance when left/right-wings share different opinions which is the most probable scenario. For instance, if the topic was started by a right-wing party, and right-wing users agreed to it including their community deceptive trolls, disagreements stances would be considered high trustworthy. If right-wing users disagreed on a topic initiated by their biased sphere and circulated by their trolls while the latter agreed on the claim, disagreements stances credibility would even become higher, regardless of how left-trolls responded to it.

B. Social Media as an Environment

Given Twitter as an example, Fig. 2 illustrates the social engagement of main tweets X_i, X_j, X_k and their relationships (1=*agree*/ -1=*disagree*) with other reactions such as other main tweets, re-tweets, replies, and pressing a love button. Since social media have a lot of uncertainty and noise, we should differentiate between two scenarios. The first case is a certain environment where stances are certain guidelines for the misinformation detection task. The second scenario is the uncertainty about such engagements, since they might be biased instead of being subjective. Also they might be manipulated by other deceptive factors such as deceptive accounts.

Stance-based methods assume that the majority of opinions would be trusted. That means the more common opinion a single main tweet X is sharing, the more likely it is not false. However, it is crucial to define that majority since a polarized political discussion has an extreme deception possibility, even for its majority of opinions. In our proposed approach, we define a more likely credible user opinion according to its unbiased measures along with other main factors such as evidence indicating a less manipulation by deceptive content and a more variety of social content the user is exposed to. Hence, all biased and immature opinions should be more likely low trustworthy.

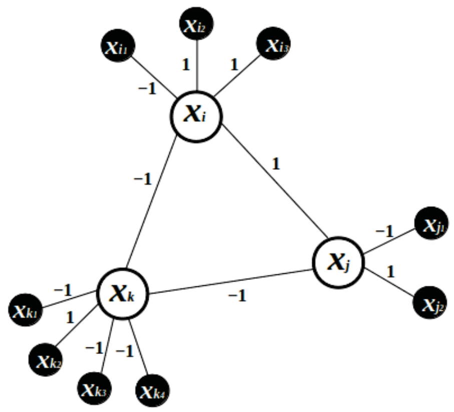

Fig. 2. Main tweets and their social engagement

Fig. 3 indicates the complexity of the problem as a noisy transformation from certainty to uncertainty. The latter occurs because the stance detection task is a probabilistic solution to opinion mining problems. Moreover, detecting a social engagement E with a probability $Pr(E)$ close to unity from a stance detection model could still be misleading, since it could be a biased opinion or influenced and manipulated by other false information or driven by psychological reasons.

Stance detection models only infer about the semantics of an opinion and do not consider how honest that opinion was.

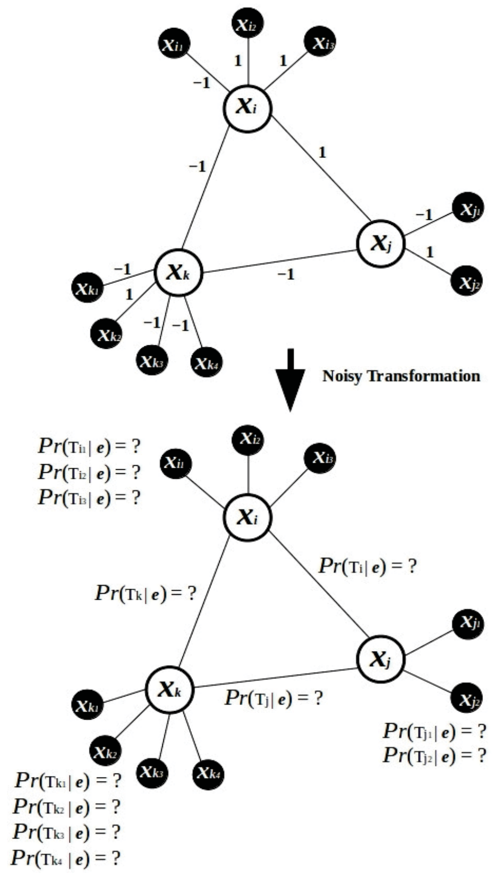

Fig. 3. Noisy transformation from certainty to uncertainty

Hence, it is more convenient to define the uncertainty of social engagement trustworthiness as a conditional probability where the veracity of the stance is depending on other factors such as polarization and other causal relations in the social network. These factors could be referred to as evidence e. Therefore, we define the social engagement trustworthiness as $Pr(T|e)$. It is useful to formalize and analyze that transformation process using a causality and a probabilistic model to represent the uncertainty following the three layers of causality: *observing*, *intervention*, and *sensitivity analysis* [19].

C. Polarization Definition

The problem of misinformation detection is highly correlated with polarization in a political crisis. Therefore, a definition of polarization is critical to the problem analysis and solution. We define an honest opinion with regard to its root causes in the network. One of these causes is the polarized community an opinion is driven by. We consider polarized opinions in a political crisis to be categorized into the following main categories which represent the possible communities in the social network: *Far left*, *left*, *neutral*, *right*, and *far right*.

Our proposed categorization should differentiate between misleading opinions and biased ones, since a false stance is always misinformation but any of the five bias levels of opinions could be either misinformation or not. Moreover, we define a polarized opinion as a relative value where one social media stance can be considered less biased with regard to another content (agree with another community claim), while the same first stance could disagree with its own community claim. In our opinion, such relative definition per each case is useful for credibility assessment, for instance, if two tweets disagreed while they belong to the same community, such stance is very important since it indicates a certain level of subjectivity and a lower polarization level.

Some literature considered the less diversity of social content as a major cause of polarization, that means the more diversity of content a user is exposed to, the less polarized the user could be in most cases [28]. Furthermore, polarization is not only influencing normal users trustworthiness, it also dictates the objectively active deceptive accounts on the network. For instance, different community deceptive accounts would try to influence their communities such as right-wing and left-wing trolls, each would try to motivate its community in a certain direction with regard to a certain topic. Typically, these directions are opposite. Hence, agreeing with a deceptive content from the right-wing would mean disagreement with another from the left-wing. Therefore, there should be some measurements for which of these biased deceptive stances are less trustworthy and which ones are ironically higher in their trustworthiness.

D. Notations

Table I describes the problem notations and their descriptions.

Definition of Bayesian Network: Let $BN = (G, \theta)$ be the Bayesian network as the pair of directed asyclic graph (DAG) G and θ as conditional probability tables (CPTs) set. Let $Z = \{T, E, P, V, D, I, L, Y, B\}$ be the set of discrete random variables (nodes) of G, where the edges are causal relations over Z.

TABLE I. PROBLEM NOTATIONS AND DESCRIPTIONS

Notations	Descriptions
(G, θ)	Bayesian network
G	Directed asyclic graph (DAG)
θ	Conditional probability table set (CPTs)
Z	Set of random variables (network structure nodes)
e	Some evidence over the network
$pa(Z)$	Parent node(s) of Z
$Y(Z)$	Child node(s) of Z
T	Trustworthiness
E	Social engagement (opinion)
P	Polarization level
V	Social content variety exposure
D	Concluded deception
L	Troll
Y	Cyborg
B	Deceptive bot
I	Topic initiative

Definition of Trustworthiness Degree: Let's denote $T = [1, 10]$ as a discrete random variable where its value ranges between 1 and 10, indicating lower to higher degree of an opinion trustworthiness, respectively. Hence, the trustworthiness degree of the ith user stance E_i can be denoted as $Pr(T_i|e)$. Where e are all the occurred evidence calculated through (G, θ) when E_i had a certain value.

Definition of Stance: Let's denote E as a discrete random variable for the network social engagements (stances) where $E = \{-1, 0, 1\}$ (*disagree*=-1/ *neutral*=0/ *agree*=1).

Definition of Polarization: Let $P = [1, 10]$ be the discrete random variable for the user polarization degree. P value ranges between 1 and 10, indicating lower to higher degree of polarization, respectively. $Pr(P_i|e)$ is the probability of the ith user polarization degree given evidence e.

Definition of Content Variety Exposure: Let $V = [1, 10]$ be the discrete random variable for the user social content variety exposure degree. V value ranges between 1 and 10, indicating lower to higher degree of content exposure, respectively. $Pr(V_i|e)$ is the probability of the ith user content exposure degree given evidence e.

Definition of Topic Initiative: Let I be the discrete random variable for the topic initiating polarized party, where $I = \{-2, -1, 0, 1, 2\}$ indicating *far left*, *left*, *neutral*, *right*, and *far right*, respectively.

Definition of Deception: Let's denote D as a discrete random variable for the concluded deceptive content opinion, where $D = \{-1, 1\}$ (*disagree*=-1/ *agree*=1). D can be observed as an evidence through its root causes, for instance, $Pr(D|D_L, D_R, I)$, where D_L, D_R, I are left/right-wings communities deceptive content stances, and the community which initiated the topic, respectively.

IV. Causal Modelling

In this section, we explain our causality-based approach to clarify the dynamics and relationships that shape the spreading of misinformation on social media during a political crisis. The conducted causality analysis in this paper should explicitly demonstrate our hypothetical assumptions on the problem as discussed in the previous sections. We aim to ask a question of which interventions are highly linked to information veracity rather than asking a prediction question only. Therefore, our main task is to model the cause and effect of the major variables on a social network that might influence or be affected by misinformation.

A causal graph is a visual representation of our assumptions about the problem and its data generating process. It should demonstrate the dynamics and relationships of the problem main entities (nodes) and the dependencies which are results of causal relations. In a causal graph, edges from parent nodes to child nodes mean a causal relationship. A child node variable is considered as an effect of its parent node variable. Fig. 4 shows a causal graph of a social network from information veracity perspective. There could be different hypothesized causal graphs for the same problem, hence, different probabilistic graphical models could be constructed as well. Evaluating different causal models is recommended in that case. In this theoretical paper, we provide one assumption of the problem and the given causal graph shows the details of this assumption.

A. *Causal Graph*

In Fig. 4, we consider a community deceptive content to be a common cause of users social engagement on that community. The assumption is that all users are distributed across different communities on social media, representing their mindset and preferences, each community will be influenced somehow by being exposed to a deceptive content targeting that community. The idea of small communities dedicated deceptive content is crucial to the challenge of biased and manipulated opinions, since by investigating such causal relations, we would be able to weight different opinions according to their causes.

In general, we consider the three main potential deceptive accounts to be the cause of a community manipulation: *trolls*, *cyborgs*, and *deceptive social bots*. In a highly polarized discussion, people would be easily manipulated and would agree on what is deceptively influencing them in their social bubbles. Hence, normal users trustworthiness degrees of each reaction to deceptive information should be measured to evaluate their credibility. Moreover, the trustworthiness degree is affected by the measures of polarization levels and social content variety exposure, the latter is directly influencing both polarization levels and social engagements as well [28]. Eventually, we consider other social engagement that might influence one's engagement like when a user is replying to others and approving or denying their opinions.

As mentioned in the problem definition, a topic initiative and a concluded deception stance from different community deceptive content should be defined in order to collect more evidence about the stances trustworthiness degrees. In our causal graph, we consider the concluded deceptive stance as a result of measuring its hypothesized causes. These causes are the community which has initiated topic, the stance on the initiated topic, and stances from other community deceptive content.

B. *Graph Semantics*

There are three main structures a causal graph could have and each one describes a unique concept of how the joint probability distribution function will be factorized. These causal semantics guide the creation of the conditional probability tables (CPTs). These CPTs are crucial since they are the model parameters. Fig. 5 demonstrates the three different causality graph structures. In the chain structure, a cause Z is influencing an effect X, the latter will trigger another effect Y. That indicates how a directly connected child node is dependent to its parent node. Moreover, that pattern holds one important property and it is crucial to the computation, which is that Y is conditionally independent from Z given that the intermediate node X occurred. By given X, we can infer about Y even if we do not know anything about Z. That conditional independence is denoted as $(Y \perp Z|X)$. That means if X occurred, $Pr(Y|Z, X) = Pr(Y|X)$ and that simplifies the calculation. In the common cause structure, Y, Z, and K are also conditionally independent if X occurred. In such causal pattern, X is called a confounder of Y, Z, and K as it is considered a common cause and they are dependent on it.

As the opposite to the previous described casual structures, the collider path or the common effect structure is different when it comes to the definition of its conditional independence,

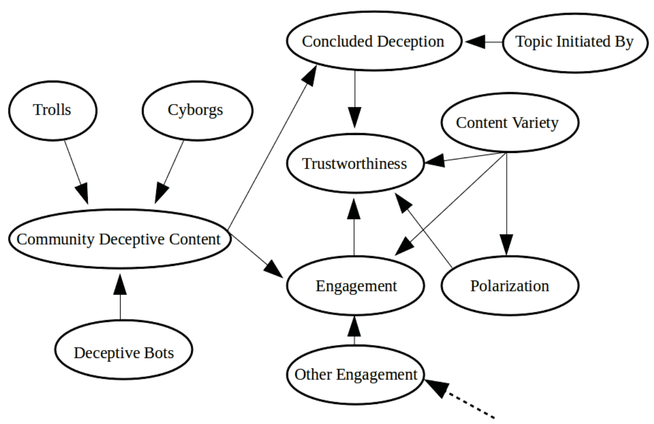

Fig. 4. Polarized political social media discussion causal graph

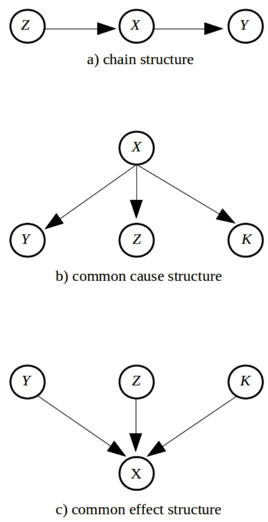

Fig. 5. Possible causal structures

so by given that X occurred, Y, Z, and K are conditionally dependent on each other which is denoted as $(Y \not\perp Z, K|X)$. A special case for the collider path is when X is a child node for another parent node $pa(X)$, if $pa(X)$ occurred, then Y, Z, and K are also conditionally dependent to each other, even if we do not know about X.

V. Methodology

A. Causal Bayesian Network

Bayesian Networks (BNs) are fundamental methods in the field of Artificial Intelligence. They provide efficient ways to calculate large and complex probabilistic inference tasks under uncertainty [32], [36], [37]. The relations in the network (directed edges) can be causal relations and the network is constructed as a directed asyclic graph (DAG) where no loops inside any part of the graph can be found. The DAG property is also important for how the reasoning would be performed, since variables independence in DAGs is compatible with how we can calculate the joint probability distribution of all the random variables. Fig. 6 shows an abstract BN, modeled according to our hypothesized causal graph (see Fig. 4) with the defined domain variables (see Table. I).

Our proposed BN is a connected graph and its complexity is bounded by the number of stances E it will investigate. Abstractly, Fig. 6 has three social engagements where E_i and E_j belong to a left-wing community, and E_k is considered a part of a right-wing community. Although E_i and E_j are social engagements from the same social bubble, they could have different polarization levels P_i, P_j if they have different social content exposure measures V_i, V_j, for instance, if user i is more exposed to other community related content.

Each stance on the given BN represents a possible scenario of interaction such as the case when E_i is influenced by its community deceptive content D_L, its exposed social content

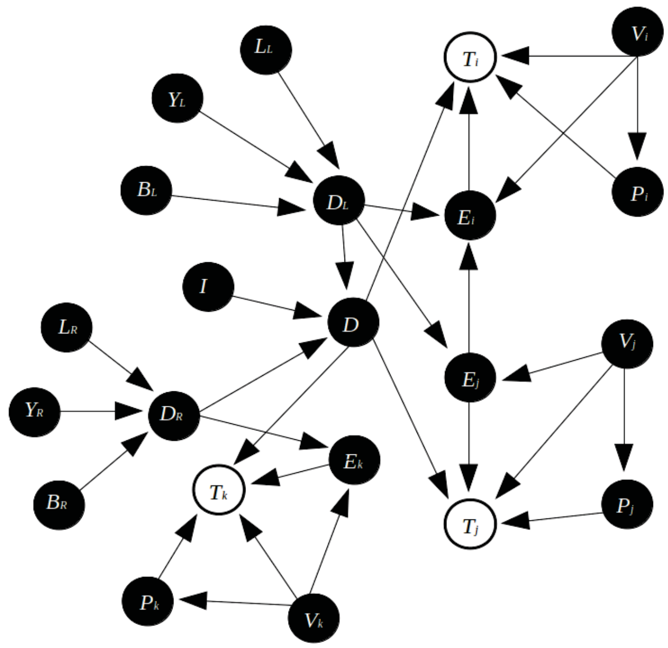

Fig. 6. The derived Bayesian Network from the assumed causal graph

variety estimation V_i, and another stance E_j that motivated user i to reply. On the other hand, E_k is only influenced by its community deceptive content D_R and its exposed content variety estimation variable V_k. In the real world scenarios, the number of nodes on such BN could be extremely bigger and the connectivity degree of the graph will be remarkably higher.

In BN, we can perform two possible types of reasoning: *predictive* and *diagnostic*, where each one dictates the direction of reasoning on the graph, either from the child node to a parent node (bottom-up) or the other way (top-down). The task of any Bayesian network is to calculate a marginal posterior probability of an unknown variable given some prior probabilities and likelihoods for other known variables. The process of calculating a marginal posterior probability is called belief update or probabilistic inference.

To build up a simulation model based on BN, we should first obtain some prior and conditional probabilities. Prior and conditional probabilities can be obtained from observations and conditional frequencies on data samples. Equation. 1 demonstrates how the joint probability distribution of all discrete random variables on BN is calculated. Furthermore, the factors of the joint probability distribution function are interpreted as CPTs for child nodes and prior probabilities for root nodes. These probabilities are considered the network parameters for calculating the targeted unknown variable. Fig. 6, declares these unknown variables with a white circle, while other black circles are representing observed evidence (assignments of variables). For example, evidence that are collected by applying stance detection, polarization estimation, and exposed content variety estimation. Moreover, deceptive accounts detection tools should be applied to collect evidence about the concluded deceptive content stance D on the network. What remains after collecting these evidence, is to calculate the marginal posterior probability of the discreet random variable T which represents the trustworthiness degree of the user social engagement.

$$Pr(z_1, ..., z_n) = \prod_{i=1}^{n} pr(z_i|pa(z_i)) \tag{1}$$

B. Belief Update

The task of the BN belief update algorithm is to learn the posterior joint probability distribution along with the network topology. There are different update belief algorithms [38]. In this section, we will give a brief statement on the EPIS-BN algorithm, which is an evidence pre-propagation importance sampling algorithm for Bayesian Networks [39]. In general, importance sampling algorithms seem to be more successful with extremely unlikely evidence, which would be the case for social media remarkable randomness. It has been stated that exact inference in Bayesian Networks is NP-hard [40]. Moreover, with thousands of variables in the network, it becomes infeasible to obtain an exact inference. Sometimes, the only way to obtain results is the approximate inference. Approximate inference is also NP-hard [41]. In general, the complexity of the computation increases if the number of parents increases for a child node, that is because the computational cost of the many entries and calculations in the CPT.

Importance sampling-based algorithms are inherited from the family of stochastic sampling algorithms [38]. The former seem to provide a more robust performance, giving the research efforts to obtain a better importance function which is crucial to the precision of the inference. Theoretically, the convergence rate of the importance sampling-based algorithms is in the order of $\frac{1}{\sqrt{n}}$, where n is the number of samples.

In general, an update belief algorithm works by determining the number of samples and initializing the prior and conditional probability tables (CPTs) of the network. According to our proposed BN in Fig. 6, for an unknown variable T_i (user i stance trustworthiness), to collect evidence e and update the beliefs for T_i, two subsets (e^+, e^-) should be defined. These subsets declare the ancestors and descendants of T_i, respectively. Then, the algorithm constructs two types of messages calculated and accumulated through e^+ and e^-: *parent to child messages* and *child to parent messages*, respectively. Fig. 7 indicates how these messages are being propagated when updating the belief of any targeted variable Z over the Bayesian network, where Z beliefs are updated through all its incoming messages. We have used the notations $pa(Z)$ and $Y(Z)$ to refer to parents and children of Z, respectively.

In a more compact form, Equation. 2 and Equation. 3 demonstrate how to calculate the incoming messages to the ith user trustworthiness degree variable T_i over the BN.

$$\pi(T_i) = \prod_{e^+} Pr(T_i|e^+_{T_i}) \tag{2}$$

$$\lambda(T_i) = \prod_{e^-} Pr(e^-_{T_i}|T_i) \tag{3}$$

Where $\pi(T_i)$ and $\lambda(T_i)$ are representing messages sent to the stance trustworthiness variable T_i from its direct causes (E_i, P_i, V_i, D), and messages sent to T_i from its effects (no effects for T_i), respectively.

Obviously, our proposed causal BN declares that the trustworthiness node T_i has no effect on any descendants, hence, $\lambda(T_i) = 1$ in that case. On the other hand, $\pi(T_i)$ can be rewritten with Equation. 4.

$$\pi(T_i) = Pr(T_i|P_i, E_i, D) \cdot \pi(P_i) \cdot \pi(E_i) \cdot \pi(D) \cdot Pr(V_i) \tag{4}$$

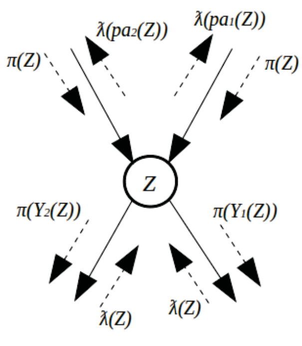

Fig. 7. Information propagation over BN

As noticed, we did not explicitly include the content variety exposure V_i variable for θ_{T_i} since it will be calculated from the messages coming from the trustworthiness node to its ancestor content variety node $\lambda(V_i)$ for the ith user. The same dropping goes for any discrete random variable that would be duplicated in the equations. Also, we have added $Pr(V_i)$ instead of $\pi(V_i)$ since content variety has no parents to receive messages from. In general, and by using Equation. 2 and Equation. 3, the belief update algorithm calculates the beliefs of a variable according to Equation 5.

$$\Pr(Z|e) = \alpha \cdot \pi(Z) \cdot \lambda(Z) \tag{5}$$

Where $\alpha = \frac{1}{e}$ as the normalization constant and the multiplication of both π and λ is a pairwise multiplication since they both are considered as probability distribution vectors over their investigated variables possible values. The result of this equation should be also a marginal posterior probability distribution for Z over the evidence e.

The importance conditional probability tables (ICPTs) are the new introduced concept to the previous general demonstrated calculations in belief update. An ICPT of a node T_i is a posterior probabilities table where $Pr(T_i|pa(T_i), e)$. The probabilities are conditional on the evidence as well, instead of conditioning only on the ancestors of T_i.

C. Toy Example

Our hypothesis about the problem of normal users credibility on polarized social media discussions is slightly tested in this section. Our toy example provides three scenarios to evaluate the proposed causal structure. We have used the EPIS-BN algorithm from GeNIe software academic version to simulate these scenarios [42]. First, we evaluate how the algorithm will perform when not all evidence are observed and the trustworthiness variable T is unknown. Second, we test the performance further by making a fully observed evidence. Third, we try to mislead the network in the second scenario by intervene and change some values for some evidence to check if there would be any contradiction in the results.

As discussed in Fig. 4 and Fig. 6, the community deceptive content is caused by objectively deceptive accounts like trolls, cyborgs, and deceptive social bots. For simplicity reasons, we have omitted the variables for these three causes and instead, we will consider only the community deceptive content variable, regardless of its causes. The main setup in the three scenarios is as follows:

- two biased communities (left-wing, right-wing) and five users are part of a political discussion: **Bob**, **Alice**, **Charlotte**, **Daisy**, and **Eric**;
- **Bob**, **Alice**, and **Eric** are part of the right-wing community, on the other hand, **Charlotte** and **Daisy** are considered members of the left-wing society;
- the community deceptive content of the left side is disagreeing on a claim, while the right side deceptive content is agreeing on it. Moreover, the topic is initiated by the right-wing community;

- the social engagement of **Charlotte** is also influenced by a social engagement from **Daisy**, and **Eric** social engagement is also influencing **Alice** opinion;

In order to initialize our proposed causal Bayesian network, CPTs should be constructed. Fig. 8, indicates an example of a CPT for the network. These values were defined as dummy data, nevertheless, they give a logical conditional frequency of how likely people would agree or disagree. In case of real data, the values could be constructed from conditional frequencies in the data itself, for example, given a time series data, how many times a user tended to agree to its own community deceptive account when the user content variety exposure was low.

Fig. 9, shows a simple community discussion over social media. Users: **Alice**, **Bob**, **Charlotte**, **Daisy**, and **Eric** were communicating with their different social background and experience. In this scenario, we have considered that not all evidence were observed and the task is to update the belief of the five users trustworthiness degree T, given the collected evidence for all causes of T except the stances E.

In Fig. 9 scenario, the discrete random variable I was indicating that the topic was initiated by the right-wing community (either normal users or deceptive accounts in that social bubble). Then, the right side deceptive accounts stances agreed on the claim of the topic, then, the left side disagreed. Since the topic was initiated by the right side and the right-wing deceptive accounts reacted with agreements, the BN updated the belief of the discrete random variable D and considered the agreement stance as the deceptive stance for the topic claim with probability $Pr(D = 1|e) = 83\%$. Since normal users stances E were not given as part of the evidence, the BN calculated their beliefs according to the incoming messages for all the corresponding nodes E_A, E_B, E_C, E_D, E_E for users **Alice**, **Bob**, **Charlotte**, **Daisy**, and **Eric**, respectively.

Notably, both **Charlotte** and **Daisy** were already a left-side community members and they both were highly polarized, hence, they both contradicted with the right-wing initiative and disagreed on it. Moreover, the more the user will disagree on the claim, the higher the trustworthiness degree will be. For instance, **Charlotte** would disagree with a belief $Pr(E_C = -1|e) = 91\%$. On the other hand, **Daisy** would disagree with a belief $Pr(E_C = -1|e) = 75\%$. Consequently, the beliefs for T_C and T_D were 61% and 55%, respectively. Furthermore, **Bob** has a higher belief of disagreement and trustworthiness $Pr(E_B = -1|e) = 98\%$, $Pr(T_B = 10|e) = 82\%$, even if he was a right-wing, that might be because of the evidence which indicated his less polarization and high exposure to diversity of content. In addition, it was noticed how **Alice** was considered less trusted since her stance belief was almost to agree and to share the same deceptive stance $Pr(E_A = 1|e) = 78\%$, $Pr(T_A = 0|e) = 74\%$.

In the second scenario, Fig. 10 explains what has happened when we replaced the beliefs probabilities of E_A, E_B, E_C, E_D, E_E with certain evidence to increase the probabilities in the first scenario to be certain values with a probability equal to unity. For instance, from $Pr(E_A = 1|e) = 78\%$ to just $E_A = 1$, and from $Pr(E_D = -1|e) = 75\%$ to just $E_D = -1$. Then, the updated beliefs of the trustworthiness degree of users became closer to 1. For instance, **Daisy** high trustworthiness degree belief changed from 55% to 65%, after giving more evidence and information. Same occurred to **Eric**, since his high trustworthiness degree in the first scenario was 51%, giving that the beliefs of his stance were distributed as $Pr(E_E = 1|e) = 47\%, Pr(E_E = 0|e) = 1\%, Pr(E_E = 1|e) = 52\%$. However, in the second scenario with a given evidence of how he has reacted exactly, his high trustworthiness degree belief became 67%.

Fig. 11 indicates an intervention in the experiment as the third case scenario. Given the first two scenarios, it was always a majority stance which was considered as a high trustworthy. For example, first two cases considered the agreement as a less trustworthy social engagement while only one out of five users had such an opinion. Our third situation tried to evaluate the challenge of a biased majority of opinions that might mislead any stance or propagation-based misinformation detection solution. Hence, we intervened to make the agreement stance as the major opinion in the discussion with even more confusing evidence, for instance, we made **Daisy** agrees but also we made her a less polarized person. Nevertheless, the results in Fig. 11 shows how the disagreement stance still considered as a high trustworthy despite of being a minority.

Bob Social Exposure	Low		Medium		High	
Deceptive Content	Agree	Disagree	Agree	Disagree	Agree	Disagree
Agree	0.67	0.14	0.47	0.42	0.01	0.98
Neutral	0.02	0.11	0.01	0.03	0.01	0.01
Disagree	0.31	0.75	0.52	0.55	0.98	0.01

Fig. 8. Bob social engagement CPT

VI. Conclusion

In this paper, we have introduced a theoretical study for the problem of normal users credibility on social media in a political crisis. Our proposed methodology could be a novel solution to the problem of misinformation. We have modeled the problem of misinformation in social media as a cause and effect process, where causes and effects are evidence to be collected before calculating the marginal posterior probability of the trustworthiness degree of the user opinion about a claim. On the other hand, recent approaches on misinformation lack the definition of polarization and biased opinions along with a full adoption to the causality approach. For instance, how traditional misinformation stance and propagation-based methods would be less efficient in polarized situations. Hence, it is crucial to define the uncertainty that occurs in a polarized political discussion over social media. Such uncertainty could not be only the extreme biased opinions as anomalies in the data, therefore, it would be more efficient to define the cause and effect between all key variables including the polarization causes, effects, and the effects of the effects. Our proposed causal Bayesian network considered these key variables as the social engagement (stance), the polarization level, the amount of information and its variety a user is exposed to, and the deceptive content in the discussion along with the topic initiative. Our toy example provided three scenarios representing partial observation of these variables, full observation, and an intervention scenario to evaluate any contradiction in the proposed causal structure.

Along with updating the beliefs for the normal users stances trustworthiness degrees, the given study would be suggested to trace the deceptive accounts, predict stances, and

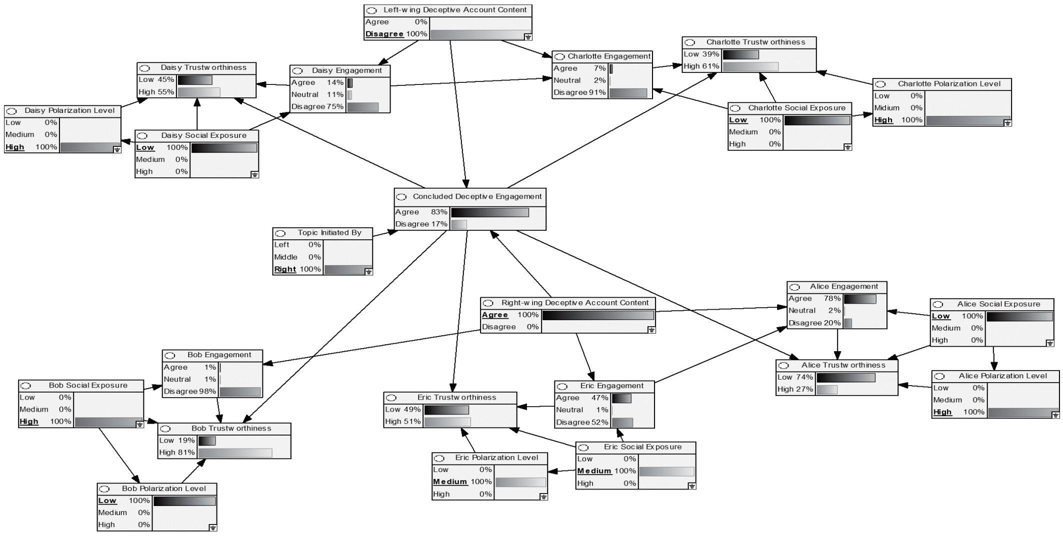

Fig. 9. Partially observed evidence scenario

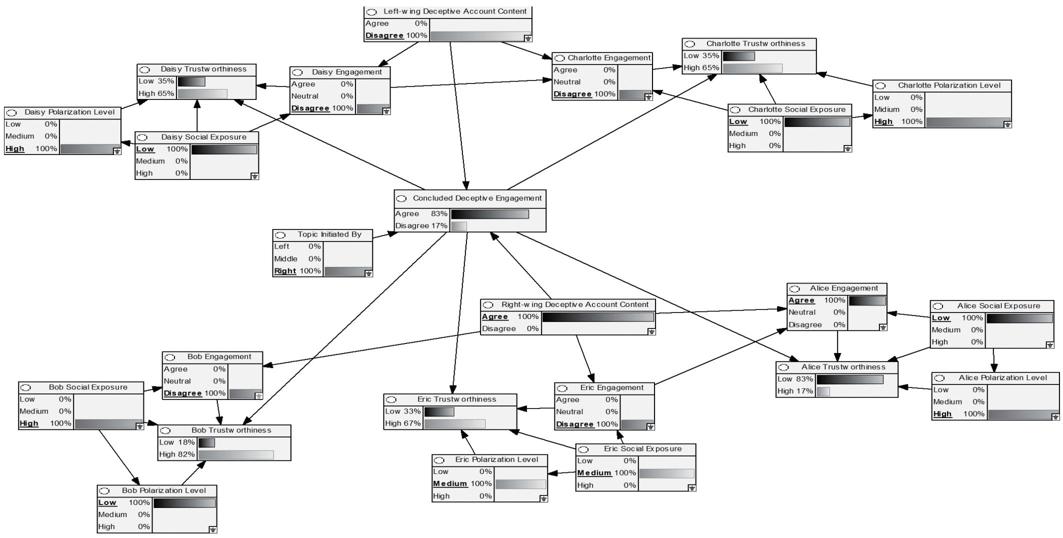

Fig. 10. Fully observed evidence scenario

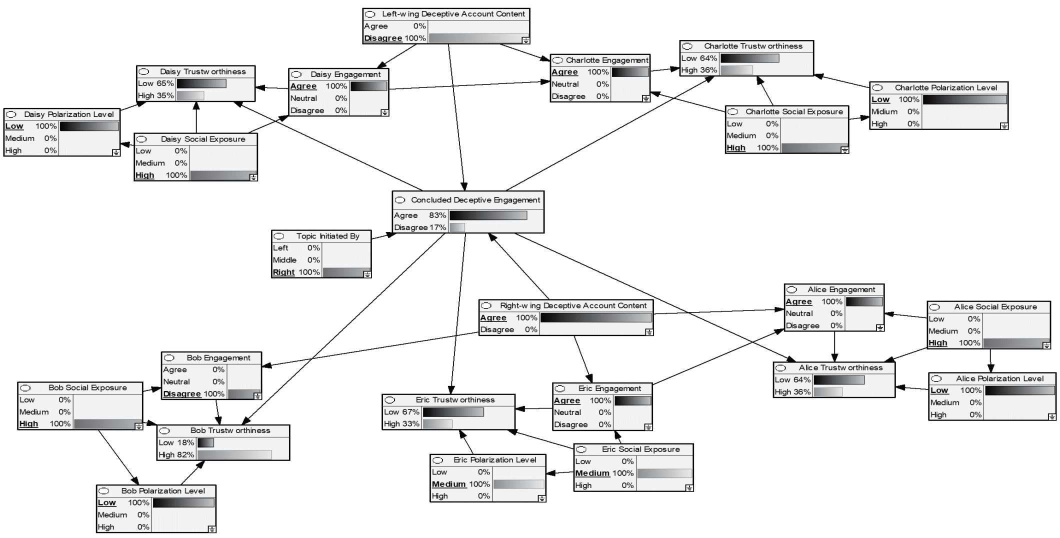

Fig. 11. Intervention scenario

estimate polarization levels. Eventually, that would lead to the computation of each normal user credibility by employing a dynamic Bayesian network DBN to infer the trustworthiness degrees of users stances over time as a temporal feature for the credibility assessment [36]. Furthermore, the proposed approach would be applied on other domains such as fake reviews on commercial products or disasters caused by natural hazards, by modeling the problem causal relations and variables within a causal Bayesian network.

In order to adopt with the complexity of the social network and the numerous number of nodes our final BN would reach, the study of how to design the system with a proper computational cost is necessary. In addition, further work should be applying some experiments based on artificial and real world data. Moreover, a complete sensitivity analysis and intervention simulation should be studied and applied on all demonstrated variables. Finally, the study of the Dynamic Bayesian Network (DBN) is important since the time dimension is critical to our problem, especially for measuring the temporal trustworthiness of normal users along with polarization and content exposure correlation.

References

[1] C. Reuter and M.-A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 41–57, 2018.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[3] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, p. 32, 2018.

[4] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web*, ACM, 2013, pp. 729–736.

[5] C. Paul and M. Matthews, "The russian "firehose of falsehood" propaganda model," *Rand Corporation*, pp. 2–7, 2016.

[6] W. Quattrociocchi, A. Scala, and C. R. Sunstein, "Echo chambers on facebook," *Available at SSRN 2795110*, 2016.

[7] N. DiFonzo and P. Bordia, "Rumor, gossip and urban legends," *Diogenes*, vol. 54, no. 1, pp. 19–35, 2007.

[8] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015.

[9] A. Bessi and E. Ferrara, "Social bots distort the 2016 us presidential election online discussion," *First Monday*, vol. 21, no. 11-7, 2016.

[10] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012.

[11] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone can become a troll: Causes of trolling behavior in online discussions," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, ACM, 2017, pp. 1217–1230.

[12] G. Bolsover and P. Howard, "Chinese computational propaganda: Automation, algorithms and the manipulation of information about chinese politics on twitter and weibo," *Information, Communication & Society*, pp. 1–18, 2018.

[13] T. Renault, "Market manipulation and suspicious stock recommendations on social media," 2017.

[14] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: Can we trust what we rt?" In *Proceedings of the first workshop on social media analytics*, ACM, 2010, pp. 71–79.

[15] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[16] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro, "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*, 2017.

[17] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[18] R. Van de Schoot, D. Kaplan, J. Denissen, J. B. Asendorpf, F. J. Neyer, and M. A. Van Aken, "A gentle introduction to bayesian analysis: Applications to developmental research," *Child development*, vol. 85, no. 3, pp. 842–860, 2014.

[19] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," *arXiv preprint arXiv:1801.04016*, 2018.

[20] A. Vlachos and S. Riedel, "Fact checking: Task definition and dataset construction," in *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, 2014, pp. 18–22.

[21] B. Shi and T. Weninger, "Discriminative predicate path mining for fact checking in knowledge graphs," *Knowledge-based systems*, vol. 104, pp. 123–133, 2016.

[22] M. Steller, "Recent developments in statement analysis," in *Credibility assessment*, Springer, 1989, pp. 135–154.

[23] W. Y. Wang, "" liar, liar pants on fire": A new benchmark dataset for fake news detection," *arXiv preprint arXiv:1705.00648*, 2017.

[24] M. Potthast, T. Gollub, M. Hagen, and B. Stein, "The clickbait challenge 2017: Towards a regression model for clickbait strength," *arXiv preprint arXiv:1812.10847*, 2018.

[25] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *2014 IEEE International Conference on Data Mining*, IEEE, 2014, pp. 230–239.

[26] M. H. Ribeiro, P. H. Calais, V. A. Almeida, and W. Meira Jr, "" everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation," *arXiv preprint arXiv:1706.05924*, 2017.

[27] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and

measuring political homophily in twitter using big data," *Journal of communication*, vol. 64, no. 2, pp. 317–332, 2014.

[28] A. Matakos and A. Gionis, "Tell me something my friends do not know: Diversity maximization in social networks," in *2018 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2018, pp. 327–336.

[29] J. Han, W. Li, and W. Deng, "Multi-resolution community detection in massive networks," *Scientific reports*, vol. 6, p. 38 998, 2016.

[30] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[31] Z. Zhao, C. Li, X. Zhang, F. Chiclana, and E. H. Viedma, "An incremental method to detect communities in dynamic evolving social networks," *Knowledge-Based Systems*, vol. 163, pp. 404–415, 2019.

[32] A. C. Constantinou, N. Fenton, W. Marsh, and L. Radlinski, "From complex questionnaire and interviewing data to intelligent bayesian network models for medical decision support," *Artificial intelligence in medicine*, vol. 67, pp. 75–93, 2016.

[33] Y. Jiang, Z. Li, and S. L. Cutter, "Social network.1667em activity space.1667em sentiment.1667em and evacuation: What can social media tell us?" *Annals of the American Association of Geographers*, pp. 1–16, 2019.

[34] O. Fraisier, G. Cabanac, Y. Pitarch, R. Besançon, and M. Boughanem, "Uncovering like-minded political communities on twitter," in *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ACM, 2017, pp. 261–264.

[35] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[36] J. Radianti, O.-C. Granmo, P. Sarshar, M. Goodwin, J. Dugdale, and J. J. Gonzalez, "A spatio-temporal probabilistic model of hazard-and crowd dynamics for evacuation planning in disasters," *Applied Intelligence*, vol. 42, no. 1, pp. 3–23, 2015.

[37] S. Glimsdal and O.-C. Granmo, "A bayesian network based solution scheme for the constrained stochastic on-line equi-partitioning problem," *Applied Intelligence*, vol. 48, no. 10, pp. 3735–3747, 2018.

[38] H. Guo and W. Hsu, "A survey of algorithms for real-time bayesian network inference," in *Join Workshop on Real Time Decision Support and Diagnosis Systems*, 2002.

[39] C. Yuan and M. J. Druzdzel, "Importance sampling algorithms for bayesian networks: Principles and performance," *Mathematical and Computer Modelling*, vol. 43, no. 9-10, pp. 1189–1207, 2006.

[40] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artificial intelligence*, vol. 42, no. 2-3, pp. 393–405, 1990.

[41] P. Dagum and M. Luby, "Approximating probabilistic inference in bayesian belief networks is np-hard," *Artificial intelligence*, vol. 60, no. 1, pp. 141–153, 1993.

[42] GeNIe software academic version, Web: https://www.bayesfusion.com/genie/.