CASINO: Towards Conformity-aware Social Influence Analysis in Online Social Networks

Hui Li
herolee@pmail.ntu.edu.sg

Sourav S Bhowmick
assourav@ntu.edu.sg

Aixin Sun
axsun@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University, Singapore, 639798

ABSTRACT

Social influence analysis in online social networks is the study of people's influence by analyzing the social interactions between individuals. There have been increasing research efforts to understand the influence propagation phenomenon due to its importance to information dissemination among others. Despite the progress achieved by state-of-the-art social influence analysis techniques, a key limitation of these techniques is that they *only* utilize *positive* interactions (*e.g.,* agreement, trust) between individuals, ignoring two equally important factors, namely, *negative* relationships (*e.g.,* distrust, disagreement) between individuals and *conformity* of people, which refers to a person's inclination to be influenced.

In this paper, we propose a novel algorithm CASINO (**C**onformity-**A**ware **S**ocial **IN**fluence c**O**mputation) to study the interplay between *influence* and *conformity* of each individual. Given a social network, CASINO first extracts a set of *topic-based subgraphs* where each subgraph depicts the social interactions associated with a specific topic. Then it optionally labels the edges (relationships) between individuals with positive or negative signs. Finally, it computes the *influence* and *conformity indices* of each individual in each *signed* topic-based subgraph. Our empirical study with several real-world social networks demonstrates superior effectiveness and accuracy of CASINO compared to state-of-the-art methods. Furthermore, we revealed several interesting characteristics of "influentials" and "conformers" in these networks.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; J.4 [**Computer Applications**]: Social and Behavior Sciences

General Terms

Algorithms, Experimentation

Keywords

Social networks, Influence, Conformity, topic-based subgraphs, Signed edge, Twitter

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

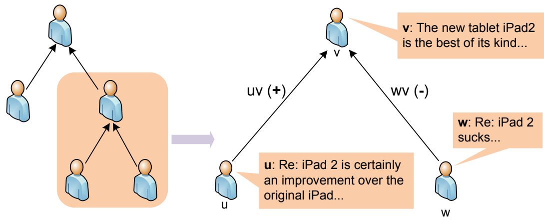


Figure 1: Social influence propagation scenario.

1. INTRODUCTION

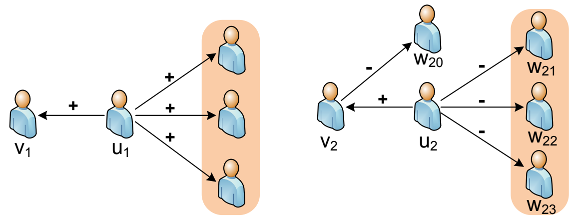
Though the field of social network analysis (SNA) has developed over many years, it is with the recent emergence of large-scale online social networking applications that techniques from this area have received a great deal of attention. We are now faced with the opportunity to analyze social network data at unprecedented levels of scale and temporal resolution. In this paper, we present our work towards addressing one of the challenges, namely the *social influence analysis* problem.

The goal of social influence analysis is to study individuals' influence by analyzing the social interactions. By identifying the "influentials" in a social network, users may be able to maximize the influence of a piece of information [6]. Informally, influentials are those individuals whose opinions or advices are often accepted and supported by others. For instance, Kempe et al. [6] proposed a greedy algorithm which aimed to find a limited number of influentials from whom the information diffusion can be maximized.

Recently, social networks are viewed as *signed* networks [7], where social interactions involve both positive and negative relationships. Consider the signed network in Figure 1 depicting interactions between a set of individuals. An edge pointing from u to v denotes that person u trust/agree (*resp.,* distrust/disagree) person v. An edge representing trust relationship is labeled as positive (*e.g.,* $\overrightarrow{uv}$), otherwise negative (*e.g.,* $\overrightarrow{wv}$). Note that social influence flows in the opposite direction of the edges (*i.e.,* v influenced u and w).

A closer analysis of social influence phenomenon in signed networks reveals that there are three important factors. Firstly, an individual's ability to influence others (*e.g.,* v). Secondly, the nature of social interactions (positive/negative) between individuals (*e.g.,* $\overrightarrow{uv}$, $\overrightarrow{wv}$). Lastly, *conformity* of an individual, which is a person's inclination to be influenced by others [2]. Note that an individual's ability to influence or conform is *context-sensitive*. For example, in Figure 1 u conforms to v's opinion on `iPad 2`. However, it does not necessarily mean that u will always conform to v on *any* topic. For instance, u may not agree with v on conversation related to `salsa dancing` as u may believe that she is a better dancer than v.

Despite the benefits of the state-of-the-art social influence analysis techniques, a key limitation is their inability to systematically



(a) (b)

Figure 2: Conformity and negative edge effect.

exploit the aforementioned second and third factors for superior analysis. Consider the two signed networks in Figure 2 where an edge $\overrightarrow{uv}$ with positive (*resp.,* negative) sign indicates u trust (*resp.,* distrust) v. The shadowed part depicts the conformity of individuals u_1 and u_2. Specifically, u_1 is easily convinced by others whereas u_2 is not. Thus, it is easier for v_1 to influence u_1 than v_2 to influence u_2. However, state-of-the-art approaches have ignored this issue and failed to differentiate between these two cases. Thus, they may compute the same *influence score* for v_1 and v_2. However, an individual's influence should be increased if a larger number of users conform to her (positive interactions) but decreased if she is distrusted by many individuals (negative interactions).

We propose a novel algorithm called CASINO (**C**onformity-**A**ware **S**ocial **IN**fluence c**O**mputation) that somewhat departs from existing influence analysis techniques in the following way: CASINO focuses on integrating the interplay of influence and conformity of individuals for social influence analysis by exploiting the positive and negative signs of edges. *To the best of our knowledge, we are the first to study the interplay of influentials and conformers with the goal of social influence analysis.* Additionally, CASINO is *context-aware*, allowing the same individual to exhibit different influence and conformity over different topics of social interactions.

Given a social network, if it is a *context-aware* one, then CASINO first extracts a set of *topic-based subgraphs*. Each subgraph depicts the social interactions between individuals associated with a specific topic. Since the edges of a social network may not always explicitly labeled with positive or negative signs, CASINO exploits an existing sentiment analysis technique to label the edges in each topic-based subgraph. Finally, given a set of signed topic-based subgraphs, the algorithm iteratively computes the *influence* and *conformity indices* of each individual in each subgraph. By applying CASINO to real-world online social media sites (*Slashdot* (`slashdot.org`), *Epinions* (`www.epinions.com`), and *Twitter*), we strongly demonstrate the effectiveness and superiority of CASINO compared to state-of-the-art approaches and at the same time reveal several interesting characteristics of influentials and conformers in these sites.

2. RELATED WORK

Leskovec et al. [7] used logistic regression to predict the signs of edges in signed networks by exploiting 7-dimensional degree features and 16-dimensional triad features. Cai et al. [3] proposed another feature (*i.e.,* influence) aside from the 7-dimensional degree features in [7]. A PageRank-like algorithm was developed to compute the influence of individual users and then use it as another feature in an SVM classifier to predict the signs of edges. In contrast, we not only study the influence of u on the sign of edge $\overrightarrow{uv}$ but also investigate the conformity of v and its effect on u's influence. Moreover, we also enable the same individual to exhibit different influence and conformity in different topics.

The notion of *conformity* originated in social psychology and is defined as yielding to perceived group pressure by copying the behavior and beliefs of others [2]. We are inspired by the conformity study in social psychology and utilize it to enhance social influence analysis in online social networks. We take into account the conformity in social influence and propose a model to evaluate the conformity in a social group.

3. INFLUENCE AND CONFORMITY

In this section, we formally introduce the notion of *influence* and *conformity* in the context of signed social networks.

Social interactions in online social networks can be either positive (*i.e.,* friendship) or negative (*i.e.,* distrust and opposition). For instance in *Epinions*, people can give both positive and negative ratings to other users. In online discussion sites such as *Slashdot*, users can tag other users as "friends" (positive) and "foes" (negative). In *Twitter*, the retweet can be a positive or a negative one. In the following discussion, we treat such social interaction as signed directed graph.

In a *signed* social network $G(V,E)$, each edge has a positive or negative sign depending on whether it expresses a positive or negative attitude from the generator of the edge to the recipient [7]. A positive sign indicates that the recipient supports the opinion of the generator whereas the negative sign represents otherwise. For example, Figure 2(b) depicts a signed social network. The positive edge $E^+ = \{\overrightarrow{u_2v_2}\}$ represents trust relationship while the negative ones ($E^- = \{\overrightarrow{w_{20}v_2}, \overrightarrow{u_2w_{21}}, \overrightarrow{u_2w_{22}}, \overrightarrow{u_2w_{23}}\}$) represent distrust relationships. Note that the signs on the edges are not always available explicitly. In networks such as *Epinions* and *Slashdot*, the sign of each edge is explicitly provided. However, in other networks such as blogosphere and *Twitter* the sign of each edge is not explicitly available. In this case, we need to preprocess the network using text mining methods to discover signs associated with the links (detailed in Section 4). Consequently, a social network $G(V,E)$ containing both positive and negative edges can be represented using a pair of graphs $G^+(V,E^+)$ and $G^-(V,E^-)$ denoting the induced graph of positive edges E^+ (trust/agreement) and negative edges E^- (distrust/disagreement), respectively.

In our approach, each individual (vertex) in a signed network is associated with a pair of *influence index* and *conformity index* to describe the power of influence and conformity. Reconsider the signed network in Figure 2(b). Intuitively, the influence of v_2 should increase as aggregated conformity of those who trust v_2 (*i.e.,* u_2) increases. On the other hand, the influence of v_2 should decrease if the aggregated conformity of those who distrust v_2 (*i.e.,* w_{20}) increases. Thus, the *influence index* of an individual should capture this interplay of influence and conformity and penalize her whenever necessary.

DEFINITION 1. ***[Influence Index]*** *Let $G^+(V,E^+)$ and $G^-(V,E^-)$ be the induced graphs of the signed social network $G(V,E)$. The influence index of vertex $v \in V$, denoted as $\Phi(v)$, is defined as follows.*

$$\Phi(v) = \sum_{\overrightarrow{uv} \in E^+} \Omega(u) - \sum_{\overrightarrow{uv} \in E^-} \Omega(u)$$

where $\Omega(u)$ represents the conformity index of vertex $u \in V$. ∎

Similarly, the *conformity index* of u_2 in Figure 2(b) depends on the influences of vertices which are trusted or distrusted by u_2. Intuitively, as the aggregated influence of those vertices which u_2 trust (*e.g.,* v_2) increases, u_2 is more inclined to conform to others. On the other hand, when the aggregated influence of vertices which u_2 distrust (*e.g.,*, w_{21}, w_{22}, w_{23}) increases, u_2 is less inclined to conform to others. This intuition is captured by *conformity index* which is defined as follows.

Algorithm 1: The CASINO algorithm.

Input: Social network $G(V,E)$
Output: the influence index $\mathbb{I}_A = (\Phi_A(u_1), \Phi_A(u_2), \ldots, \Phi_A(u_\ell))$ and conformity index $\mathbb{C}_A = (\Omega_A(u_1), \Omega_A(u_2), \ldots, \Omega_A(u_\ell))$ for $V = \{u_1, u_2, \ldots, u_\ell\}$ and for each topic A

1 **begin**
2   **if** *G is context-aware* **then**
3     $\mathcal{G} \leftarrow$ **extractSubgraph**(G);
4   **else**
5     $\mathcal{G} = \{G\}$;
6   **foreach** $G_A \in \mathcal{G}$ **do**
7   **if** *G_A is not a signed network* **then**
8       $(G_A^+(V_A, E_A^+), G_A^-(V_A, E_A^-)) \leftarrow$ **edgeLabel**(G_A);
9     $(\mathbb{I}_A, \mathbb{C}_A) \leftarrow$ **indicesCompute**($G_A^+(V_A, E_A^+), G_A^-(V_A, E_A^-)$);

Algorithm 2: The *edgeLabel* procedure.

Input: Topic-based subgraph $G_A(V_A, E_A)$ induced by topic A,
Output: $G_A^+(V_A, E_A^+)$ and $G_A^-(V_A, E_A^-)$ such that: $E_A^+ \cup E_A^- = E_A$ and $E_A^+ \cap E_A^- = \emptyset$

1 **begin**
2   $E_A^+ = E_A^- = \emptyset$;
3   **foreach** $\overrightarrow{uAv} \in E_A$ **do**
4     $u.sentiment \leftarrow LingPipe.$**sentExtr**($u$);
5     $v.sentiment \leftarrow LingPipe.$**sentExtr**($v$);
6   **if** $|u.sentiment - v.sentiment| < \varepsilon$ **then**
7       $E_A^+ = E_A^+ \cup \{\overrightarrow{uAv}\}$
8   **else**
9       $E_A^- = E_A^- \cup \{\overrightarrow{uAv}\}$

DEFINITION 2. ***[Conformity Index]*** *Let $G^+(V,E^+)$ and $G^-(V,E^-)$ be the induced graphs of the signed social network $G(V,E)$. The conformity index of vertex $u \in V$, denoted as $\Omega(u)$, is defined as follows.*

$$\Omega(u) \quad = \quad \sum_{\overrightarrow{uv} \in E^+} \Phi(v) - \sum_{\overrightarrow{uv} \in E^-} \Phi(v)$$

where $\Phi(v)$ is the influence index of vertex $v \in V$. ∎

Thus, according to the above definition the influence index of v_2 in Figure 2(b) can be computed as $\Phi(v_2) = \Omega(u_2) - \Omega(w_{20})$. The conformity index of u_2 is computed as $\Omega(u_2) = \Phi(v_2) - \Phi(w_{21}) - \Phi(w_{22}) - \Phi(w_{23})$. Observe that the aforementioned definitions of influence and conformity are mutually dependent on each other. Consequently, a recursive computation framework is necessary to compute these two indices.

4. THE CASINO ALGORITHM

In this section, we formally describe the algorithm CASINO for computing conformity and influence indices of individuals in a social network containing positive and negative edges. Online social networks can be classified into *context-aware* and *context-free* networks. The former represent networks where the edges are associated with topics (context) as social interactions may often involve conversations on specific topics. For example, each conversation in *Twitter* is based on a specific topic. On the other hand, interactions in context-free networks do not involve specific topics. For example, in *Epinions* and *Slashdot* individuals trust (distrust) each other regardless of any specific topic. The leftmost social network in Figure 3 is an example of context-aware social network where an edge labeled as A_1, A_2 indicates that the pair of individuals communicate with each other on topics A_1 and A_2.

The CASINO algorithm is outlined in Algorithm 1 and consists of three phases, namely the *topic-based subgraph extraction* phase (Line 3), the *edge labeling* phase (Line 8), and the *indices computation* phase (Line 9).

Figure 3 depicts an overview of CASINO algorithm. Given a social network $G(V,E)$, if it is context-aware then the *topic-based subgraph extraction* phase extracts a set of subgraphs $\mathcal{G}$ where each subgraph $G_A(V_A, E_A) \in \mathcal{G}$ contains all the vertices and edges in G associated with a specific topic A. Each subgraph G_A represents positive or negative attitudes of individuals toward opinions of others in G with respect to the topic A. For instance, in Figure 3, this phase generates three topic-based subgraphs, namely, G_{A_1}, G_{A_2}, and G_{A_3}, for topics A_1, A_2, and A_3, respectively. Recall that edges of a social network may not be explicitly labeled with positive or negative signs. This is especially true for context-aware networks (*e.g., Twitter*). On the other hand, links in many context-free networks (*e.g., Slashdot* and *Epinions*) are explicitly labeled with signs. Hence, it is important to label the edges in each topic-based subgraph G_A. The objective of the *edge labeling* phase is to assign sign to each edge by analyzing the sentiment expressed by the edge. Figure 3 depicts the labeling of G_{A_1}. Finally, given a set of signed topic-based subgraphs $\mathcal{G}$, the goal of the *indices computation* phase is to iteratively compute the influence and conformity indices of each individual in each $G_A \in \mathcal{G}$. Observe that a vertex v in G may have multiple pairs of indices if v is involved in more than one topic-based subgraph. Since the first phase is straightforward, we now elaborate on the remaining two phases in turn.

The edge labeling phase. The edge labeling method varies with dataset. In this paper, we adopt the method described in Algorithm 2. For each edge $\overrightarrow{uAv}$ (the edge pointing from u to v on context topic A) in G_A, we identify 5-level sentiment (*i.e.,* like, somewhat like, neutral, somewhat dislike, dislike) expressed at both ends using *LingPipe* [1] which is a popular sentiment mining package adopted in several recent research [4] (Lines 4-5). If the sentiments at both ends are similar (*sentiment similarity threshold* is less than ε), we denote the edge as positive (Lines 6-7). Otherwise, we denote it as negative (Lines 8-9).

Indices computation phase. Given a topic A and topic-based subgraph G_A, the preceding phase generates G_A^+ and G_A^-. Without loss of generality, assume that there are $|\mathcal{G}|$ different topics. Then, we are able to compute an individual's influence and conformity indices for each topic (*i.e.,* $\Phi_A(u)$ and $\Omega_A(u)$). Algorithm 3 outlines the strategy for computing a pair of influence and conformity indices ($\Phi(u)$, $\Omega(u)$) for each vertex u. It first initializes the influence index and conformity index of all vertices to be 1 (Lines 1-4). Subsequently, in each iteration it computes them for each vertex by using the values of the indices in previous iteration (Lines 6-8) and normalizing these values using the square root of the summation of all vertices' index values (Lines 9-13). The algorithm terminates when both indices converge. In [8], we have proved that CASINO is guaranteed to converge after a fixed number of iterations n. Observe that the technique can easily be extended to compute the aggregated indices of an individual by taking into account the entire social network G over all topics $A = 1, \ldots, |\mathcal{G}|$. In this case, E^+ and E^- in Definitions 1 and 2 are replaced by $\bigcup_{A=1}^{|\mathcal{G}|} E_A^+$ and $\bigcup_{A=1}^{|\mathcal{G}|} E_A^-$, respectively.

5. EXPERIMENTAL STUDY

In this section, we present the experiments conducted to evaluate the performance of CASINO and report some of the results obtained. A more detailed results is available in [8]. To this end, we borrow

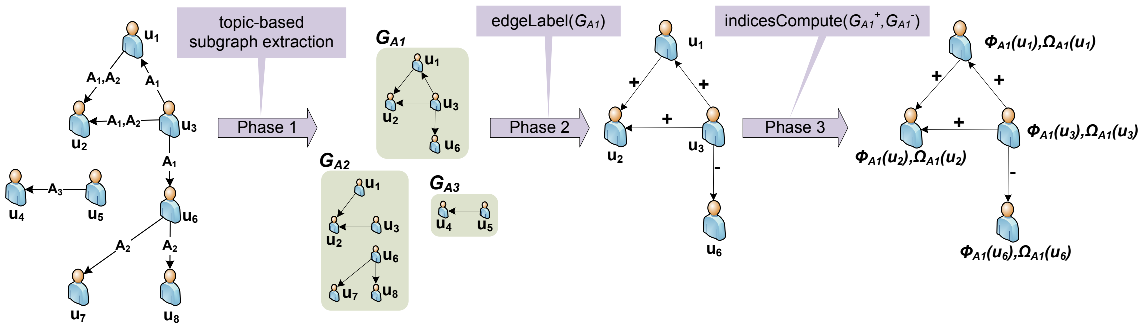


Figure 3: Overview of CASINO.

Algorithm 3: The *indicesCompute* procedure.

Input: $G(V,E) = G^+(V,E^+) \cup G^-(V,E^-)$
Output: the influence index $\mathbb{I} = (\Phi(u_1), \Phi(u_2), \ldots, \Phi(u_\ell))$ and conformity index $\mathbb{C} = (\Omega(u_1), \Omega(u_2), \ldots, \Omega(u_\ell))$ for $V = \{u_1, u_2, \ldots, u_\ell\}$

begin
  $k = 1$ /*initialize iteration counter*/ ;
  foreach $u \in V$ **do**
    $\Phi^k(u) = \Omega^k(u) = 1$
  while $\mathbb{I}$ *or* $\mathbb{C}$ *not converged* **do**
    foreach $u \in V$ **do**
      $\Phi_0^{k+1}(u) = \sum_{\overrightarrow{vu} \in E^+} \Omega^k(v) - \sum_{\overrightarrow{vu} \in E^-} \Omega(v)$;
      $\Omega_0^{k+1}(u) = \sum_{\overrightarrow{uv} \in E^+} \Phi^k(v) - \sum_{\overrightarrow{uv} \in E^-} \Phi(v)$;
    foreach $u \in V$ **do**
      $\Phi^{k+1}(u) = \frac{\Phi_0^{k+1}(u)}{\sqrt{\sum_{v \in V} \Phi_0^{k+1}(v)^2}}$;
      $\Omega^{k+1}(u) = \frac{\Omega_0^{k+1}(u)}{\sqrt{\sum_{v \in V} \Omega_0^{k+1}(v)^2}}$;
    $\mathbb{I}^{k+1} = (\Phi^{k+1}(u_1), \Phi^{k+1}(u_2), \ldots, \Phi^{k+1}(u_l))$;
    $\mathbb{C}^{k+1} = (\Omega^{k+1}(u_1), \Omega^{k+1}(u_2), \ldots, \Omega^{k+1}(u_l))$;
    $k = k + 1$;

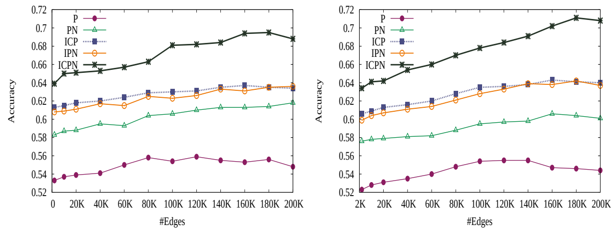
Table 1: Statistics of datasets.

Dataset	*#nodes*	*#edges*	*#positive edges*	*#negative edges*
Slashdot	77,357	516,575	396,378	120,197
Epinions	131,828	841,372	717,667	123,705
Twitter	576,894	1,230,748	1,015,492	195,256

the experimental framework articulated by Leskovec et al. [7], in which machine-learning based approach is used for discovering the presence of unknown edges and edge sign prediction by exploiting various edge features. By considering conformity and influence indices as new features for classification, it is expected that these additional features will provide more concrete evidence for edge prediction compared to state-of-the-art strategies.

We consider the following context-free and context-aware networks where each link is explicitly or implicitly labeled as positive or negative. (a) **Context-free network data:** In order to compare the performance of CASINO with state-of-the-art efforts on influence evaluation in signed networks [3, 7], we adopt the same datasets that have been used in these work: *Slashdot* and *Epinions* [1] (Table 1). Recall that these network data are context-free and contain explicit signs of edges to indicate the attitudes of individ-

[1] http://snap.stanford.edu/data/



(a) Epinions (b) Slashdot

Figure 4: Positive edge presence prediction.

uals towards one another. (b) **Context-aware network data:** We use the *Twitter* dataset to investigate the performance on a context-aware network. The dataset was crawled using the *Twitter* API [2] during Dec 2010 to Feb 2011 (Table 1). We extracted top 20 trends keywords at hourly duration and retrieved up to 1500 tweets for each trend. Then we identified the relationships between all the tweets in the dataset. The reader may refer to [8] for details related to statistics of the dataset and construction of context-aware interaction graphs from it.

5.1 Experimental Setup

We have implemented CASINO in Java and run all the experiments on a 1.86GHz Intel 6300 machine with 4GB RAM with Windows XP. The algorithm converges after average 30 iterations. We set the *sentiment similarity threshold* ε (Algorithm 2) to 1.

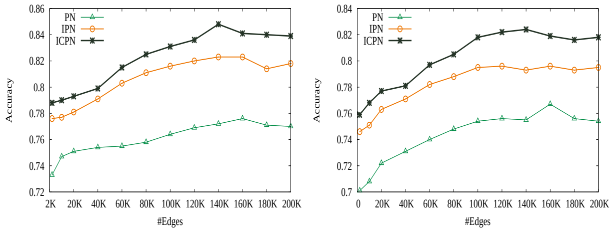
Similar to [3,7], in our experiments, the task of edge sign prediction is considered as a binary classification problem. We adopted SVM^{light} classifier [5] and classification *accuracy* is taken as the main measure for evaluation. The reason for adopting accuracy as evaluation measure over precision and recall is that the former is suitable for quantifying the prediction performances of both positive and negative samples. This is crucial in our framework as in most experiments both positive and negative edges are being predicted. In order to avoid biased classification, we adopt the same strategy used in [3, 7]. Specifically, we create a balanced dataset with same number of positive and negative edges for training and testing. Based on the learned model (discussed below), we predict a label 1 or -1 for each target edge indicating its possibility to be positive or negative, respectively.

In order to compare the prediction accuracy of the proposed approach with state-of-the-art efforts, a baseline classifier is constructed by referring to the structural features discussed in [7]. Specifically, given an edge from vertex u to v, a 7-dimensional feature vector $\{d_{in}^+(v),\ d_{in}^-(v),\ d_{out}^+(u),\ d_{out}^-(u),\ d_{in}(v),\ d_{out}(u),\ C(u,v)\}$ is constructed, where features $d_{in}^+(v)$ and $d_{in}^-(v)$ denote the number of positive and negative incoming edges to v, features $d_{out}^+(u)$ and $d_{out}^-(u)$ represent the number of positive and negative outgo-

[2] http://dev.twitter.com/doc

Table 2: Features involved in different approaches. (P: positive, N: negative, I: influence, C: conformity, A: topic)

Features \ Approaches	P	PN	IPN	ICP	ICPN	ICAPN
$d_{in}^{+}(v)$	✓	✓	✓	✓	✓	✓
$d_{in}^{-}(v)$	-	✓	✓	-	✓	✓
$d_{out}^{+}(u)$	✓	✓	✓	✓	✓	✓
$d_{out}^{-}(u)$	-	✓	✓	-	✓	✓
$d_{in}(v)$	-	✓	✓	-	✓	✓
$d_{out}(u)$	-	✓	✓	-	✓	✓
$C(u,v)$	✓	✓	✓	✓	✓	✓
$\Phi(u)$	-	-	✓	-	-	-
$\Phi(v)$	-	-	✓	✓	✓	✓
$\Omega(u)$	-	-	-	✓	✓	✓
context A	-	-	-	-	-	✓
$\Phi_A(v)$	-	-	-	-	-	✓
$\Omega_A(u)$	-	-	-	-	-	✓



(a) Epinions (b) Slashdot

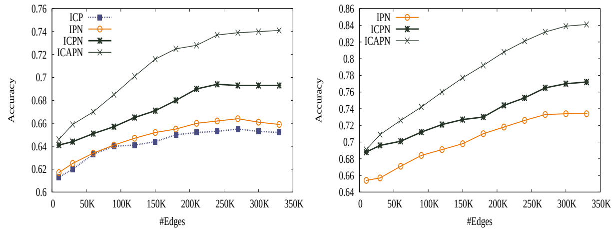
Figure 5: Signed edge prediction.

ing edges from u, features $d_{in}(v)$ and $d_{out}(u)$ denote the number of in-degree and out-degree of v and u, and $C(u,v)$ denotes the total number of common neighbors of u and v without considering the edge direction. Then, we create variants of this baseline classifier by adding influence index and conformity index as new features. Table 2 describes the features involved in the feature vector for each edge $\overrightarrow{uv}$ for different variations of the baseline classifier.

5.2 Experimental Results

Positive edge presence prediction. We first conduct a series of experiments to predict the presence of positive edges. In order to test the effect of negative edges in predicting the presence of positive edges, we adopted a regression function of SVM^{light} where presence of positive edges are labeled as 1 and negative edges are labeled as -1. Figure 4 shows the average accuracy for the benchmark datasets, where the results are compared between different classifiers for different size of training and testing data. We can make the following observations. First, for each dataset PN performs better than P indicating that information related to negative edges enhance the quality of edge presence prediction. Second, the prediction accuracy is further improved when we incorporate the influence index or conformity index as a feature (ICP and IPN). Note that ICP performs slightly better than IPN but the improvement is not significant. When the training set is large both approaches exhibit similar performance. Third, ICPN consistently reports the best prediction performance. That is, edge presence prediction is enhanced when we consider the interplay between influentials and conformers.

Signed edge prediction. Next, we undertake a series of experiments to predict the signs of edges. Our goal is to predict the signs of edges which maybe either positive or negative. Figure 5 reports the prediction accuracies of the classifiers. Observe that among the three approaches involving both positive and negative edges (PN, IPN, and ICPN), ICPN performs the best, followed by IPN and PN, respectively. Thus, by taking into account the influence and conformity of vertices, the accuracy of sign prediction task can improve significantly.



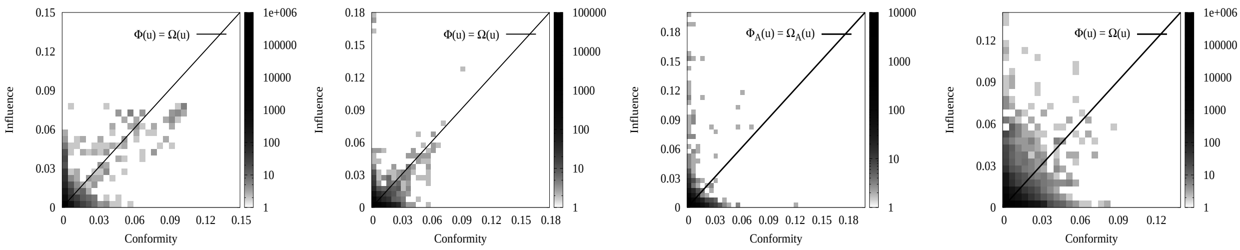
(a) Positive edge presence (b) Edge sign prediction

Figure 6: Context-aware prediction accuracy.

Edge presence and signed edge prediction in context-aware networks. We now report the performance of our model on a context-aware network (*Twitter*). In this experiment we adopt the classifier ICAPN which takes into account the topic information associated with each edge. Figure 6 plots the prediction accuracies of the relevant classifiers. Observe that in both figures ICAPN outperforms the rest. Note that the performances of ICAPN and ICPN are similar when the training set is very small. This is because there may not be enough training edges in each topic-based subgraphs G_A when the training set is very small. Consequently, not enough information is available to accurately compute $\Phi_A(v)$ and $\Omega_A(u)$. All these evidences demonstrate that by leveraging on the influence and conformity indices in topic-based subgraphs, the proposed model leads to superior prediction performance for both positive edge presence and edge sign prediction tasks.

Influentials and conformers. We analyze the list of influentials and conformers detected by the CASINO algorithm. Figure 7 depicts the distribution heatmaps of influence index versus conformity index for each benchmark dataset. For each individual u in a network we compute $\Phi(u)$ and $\Omega(u)$ and represent it as a point in the influence-conformity 2-D plane. Then we separate the plane into grids of size 0.005×0.005 and count the number of points in each grid. The color shade of a grid denotes the number of points residing in it. Note that both influence index and conformity index are normalized into the range of $[0,1)$. For each figure, we explicitly draw a boundary line along which the vertices exhibit identical influence index and conformity index. Observe that the line separates the influence-conformity plane into two areas. In the sequel, we refer to the top area as '*Area I*' and the down one as '*Area II*'. The points belonging to '*Area I*' exhibit higher influence index compared to conformity index, indicating that individuals in this area are more prone to influence others than being influenced. We refer to them as *influence-biased*. On the other hand, '*Area II*' represent individuals who are more prone to be influenced than influencing others. We refer to these individuals as *conformity-biased*.

We first analyze the context-free networks (Figures 7(a)(b)). Consider Figure 7(a) related to *Epinions* dataset. Observe that 31% of all individuals belong to '*Area I*'. Consequently, fewer number of individuals in this network are influence-biased. That is, majority of individuals in *Epinions* are often conforming to the others. Similar phenomenon also exists in the *Slashdot* dataset. Observe that those vertices in *Epinions* which exhibit very high conformity index values also have high influence index values. On the other hand, vertices with highest influence index values have a wider range of conformity indices (*i.e.,* from 0 to 0.11). Such phenomenon indicates that in *Epinions* most influence-biased individuals may also conform to others whereas the most conformity-biased individuals are always influencing others. Interestingly, the phenomenon is different in *Slashdot* dataset (Figure 7(b)). Specifically, individuals who are associated with highest influence index values have very



(a) Epinions (b) Slashdot (c) Twitter (Top-1) (d) Twitter (All)

Figure 7: Influence vs. conformity distribution heatmap.

Table 3: Top 10 authors with the highest influence index and conformity index.

Rank	Influential twitter (#positive in-links/#in-links)			Conformer twitter (#positive out-links/#out-links)		
	All	*Top-1*	*Top-2*	*All*	*Top-1*	*Top-2*
1	142987924 (66/73)	3453454 (13/13)	950596 (31/34)	51389816 (35/37)	121836131 (14/16)	49276778 (101/144)
2	49276778 (61/82)	56068621 (11/11)	190108655 (11/11)	172039151 (31/34)	105332925 (13/14)	202346609 (45/61)
3	119394881 (60/77)	3984874 (10/10)	3498571 (8/9)	177173204 (30/35)	177255919 (11/12)	197538544 (26/30)
4	231134989 (55/71)	133282617 (11/11)	147327886 (5/5)	143062806 (27/34)	193206052 (11/12)	184930795 (22/26)
5	2109823 (56/72)	199855121 (7/7)	49126931 (5/5)	128118710 (25/33)	36525648 (9/10)	148335502 (21/23)
6	92503401 (55/78)	8234375 (5/5)	121158546 (5/5)	130414633 (30/41)	90723076 (7/8)	171387567 (17/20)
7	206661373 (51/66)	1465130 (3/3)	129009252 (5/5)	4782790 (23/30)	123606641 (6/8)	126407259 (18/22)
8	220490093 (46/60)	2894822 (3/3)	79897503 (4/4)	125551983 (22/34)	51513825 (6/6)	114455733 (14/20)
9	168175236 (40/51)	21755211 (2/2)	83629945 (4/4)	91930055 (21/28)	203774695 (5/6)	217826740 (15/20)
10	171287044 (41/62)	4051581 (2/2)	166830172 (4/4)	145339829 (22/31)	203780314 (4/4)	159724683 (12/17)

small conformity index (*i.e.,* less than 0.02). But individuals with highest conformity index may not exhibit small influence index values. In fact, the conformity index values of these conformity-biased individuals are distributed along the boundary line ($\Phi(u) = \Omega(u)$). Thus, we can make the following observations regarding *Slashdot*. Firstly, there are a few influence-biased individuals who exhibit very high influence but are not easily influenced by others. Secondly, there do not exist conformity-biased individuals who are not influencing others at all.

Next, we analyze the context-aware network (Figures 7(c)(d)). Figures 7(c) show the distributions of influence and conformity indices for the top-1 topic (`Mumford & Sons`) with the most number of tweets (4390). Observe that 40% of all the individuals fall in '*Area I*' for Figure 7(c). Notably, it exhibits *certain* influence-biased characteristics similar to *Slashdot*. That is, there are a few influence-biased individuals who exhibit very high influence but are not easily influenced by others. This similarity may be due to the fact that both *Slashdot* and *Twitter* are driven by user conversations where majority individuals are commenting or following a few individuals who started the conversations. Figure 7(d) plots the distribution of indices computed over *all* topics. In this case, 41% of all individuals belong to '*Area I*'.

Table 3 shows IDs of top-10 authors who exhibit the highest influence index and conformity index for top-1, top-2 and all topics. Consider the top two twitters for all topics. The author `'142987924'` who has the highest influence index receives 66 conforming edges out of 73 in-links over 22 topics. On the other hand, the author `'51389816'` who exhibits the highest conformity index initiates 35 conforming edges out of 37 out-links over 37 topics indicating that she has high chance to conform to others' opinions in almost all the topics she is involved in. We can make the following observations. Firstly, none of the top-10 authors occupies a position in *both* indices. Secondly, the top-10 individuals having highest influence and conformity indices are different for different topics. This confirms our hypothesis that social influence phenomenon is context-sensitive as same individual may exhibit different influence and conformity over different topics of social interactions.

6. CONCLUSIONS

The social influence analysis for online social networks is an important problem with applications to viral marketing and information dissemination among others. State-of-the-art social influence analysis techniques ignore two equally important factors, namely, negative relationships and conformity of people. In this paper, we propose a novel algorithm for social influence analysis called CASINO, which quantifies the influence and conformity of each individual in a network by utilizing the positive and negative relationships between individuals. Our exhaustive experimental study using several online social media sites demonstrates the effectiveness and superior accuracy of CASINO compared to state-of-the-art methods. Specifically, our investigation revealed that the knowledge of conformity of individuals enhance the accuracy of social influence analysis. We also observed several interesting characteristics of influentials and conformers in *Slashdot*, *Epinions*, and *Twitter*.

Acknowledgement: The work of the third author is supported by MOE AcRF Tier-1 Grant RG13/10, Singapore.

7. REFERENCES

[1] A-lias. Lingpipe 4.0.1. http://alias-i.com/lingpipe, 2008.

[2] E. Aronson, T. D. Wilson, and R. M. Akert. *Social Psychology*. Prentice Hall, fifth edition, Feb. 2004.

[3] K. Cai, S. Bao, Z. Yang, et al. Oolam: an opinion oriented link analysis model for influence persona discovery. In *WSDM*, 2011.

[4] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM*, 2011.

[5] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, USA, 1999.

[6] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[7] J. Leskovec, D. P. Huttenlocher, and J. M. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, 2010.

[8] H. Li, S. S. Bhowmick, and A. Sun. CASINO: Towards conformity-aware social influence analysis in online social networks. Technical report, Available at `http://www.cais.ntu.edu.sg/~assourav/TechReports/CASINO-TR.pdf`, July 2011.