

國立臺灣大學管理學院資訊管理學系

博士論文

Department of Information Management

College of Management

National Taiwan University

Doctoral Dissertation



主題人物立場分析研究

A Study of Topic Person Stance Analysis

陳仲詠

Zhong-Yong Chen

指導教授：陳建錦 博士

Advisor: Chien Chin Chen, Ph.D.

中華民國 105 年 6 月

June 2016

國立臺灣大學博士學位論文  
口試委員會審定書



主題人物立場分析研究

A Study of Topic Person Stance Analysis

本論文係陳仲詠君（學號 D98725003）在國立臺灣大學資訊管理學系、所完成之博士學位論文，於民國 105 年 6 月 6 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

陳建錦

張嘉惠

陳通彰

蔡銘偉

陳信平

李益坤

所長：

## 謝辭

首先，感謝博士學位口試委員們陳建錦老師、陳信希老師、陳孟彰老師、張嘉惠老師與蔡銘峰老師的寶貴意見，讓此論文更臻完善。其中特別感謝我的指導教授陳建錦老師，不論在修課、討論研究方法、尋找研究主題、編修論文、甚或是做人處事的道理，老師都非常用心在指導。從老師身上我學到了許多，如此良師，甚為少見。

我也感謝周圍一起努力的學長姐(健誠、蕭鉢、澤龍)、同學(富丞、詠淳)、學弟妹(澤翰、小璧、小麥、小美、Alicia、小力、容萱、楊衡、卡布、呂芃、燕秋、李孟、施舜元、黃雅歆、陳世穎、楊智幃)，有你(妳)們，讓我的求學生涯多采多姿，有你(妳)們，也讓我的博士旅途不寂寞。我永遠忘不了跟學長們一人一間博士班實驗室的時光，我永遠忘不了跟富丞一起拼命開吃大餐的日子，我永遠也忘不了 WEAL 的歡笑、實驗室聚餐與各式各樣由你(妳)們主辦的活動。

在這條追求學問的漫長旅途上，我非常感激我的家人與我的未婚妻憶萱在我的身邊扶持我，幫我打氣，你(妳)們是我的精神支柱，有你(妳)們我才有勇氣一直走下去。

最後，此博士論文獻給所有在我博士生涯中指導我、協助我、鼓勵我與支持我的所有人。

## 中文摘要

隨著網路的爆炸性成長，人們能夠輕易地從網路獲得龐大的資訊，而且人們可能會被網路上的多媒體所提供的資訊所掩沒，像是新聞、網路評論、論壇文章或是從社群媒體來的資訊。為了協助人們消化這些資訊，在此博士論文中，我們探究一個新穎的主題，稱為主題人物立場辨識，這個主題的目的地是辨識主題文件中人物的立場。我們提出了兩套方法來解決這個問題。首先，我們提出一套叫做模式基礎 EM 的方法，利用人物名字共同出現在文件中的模式來辨識主題人物的立場。此外，文件中人名共同出現與不共同出現的程度被考量用以加權人名共同出現的模式。甚至，我們發展一個初始化演算法來穩定辨識人物立場社群，這是因為模式基礎的 EM 方法對於初始化是頗敏感的。第二套方法稱做使用友誼網路分析的主題人物立場社群辨識，這套方法考量文件的友善(敵對)傾向自動地從主題文件建構友誼網路。此外，我們提出立場擴展與立場修正演算法基於友誼網路來辨識立場社群。實驗結果驗證這兩套方法都比過去知名的分群演算法效能來得好。

關鍵字: 主題人物立場分析、主題人物分群、文字探勘、資訊檢索、分群演算法

## Abstract



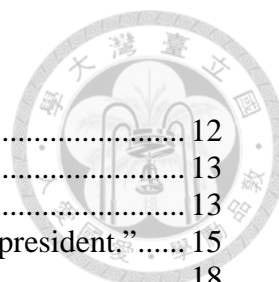
With the explosive growth of the Internet, people can easily receive astronomical information from the Web, and could be overwhelmed by the online medium, e.g. news, review comments, forum posts or information from the social medium. For facilitating the people digest the enormous information, we investigate a novel problem named “topic person stance identification,” which is to identify the stances of the topic persons from topic documents, in this dissertation. We proposed two methodologies to copy with the problem. First, we proposed a methodology named model-based EM method to identify the stances of the topic persons by leveraging the pattern of person name co-occurrence in the documents. In addition, the level of co-occurrence and non-co-occurrence of the person names in the documents are considered to weight the pattern of the person name co-occurrence. Moreover, we developed an initialization algorithm to stable the results of identifying the stance communities because the EM method is sensitive to the initialization. The second methodology is called stance community identification of topic persons using friendship network analysis. This method is to take the friendly (opposing) orientation of the documents into consideration to construct the friendship network automatically from the topic documents. For identifying the stance community, we proposed stance community expansion and stance community refinement algorithms to identify the stance communities based on the network. The experimental results of two methodologies demonstrated our methods are outperformed other well-known clustering approach, and can effectively identify the stances of the topic persons

Keywords: topic person stance analysis, topic person clustering, text mining, information retrieval, clustering algorithm

# Table of Contents

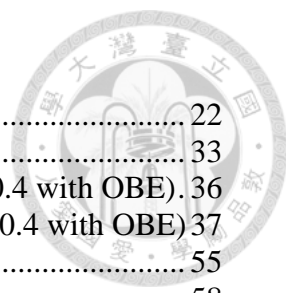


口試委員會審定書 .....	i
謝辭 .....	ii
中文摘要 .....	iii
Abstract.....	iv
1. Introduction .....	1
2. Literature review.....	4
2.1 Opinion mining.....	4
2.2 Community detection .....	6
2.2.1 Eigen-based community detection approach.....	6
2.2.2 Iterative clustering approach .....	8
3. A model-based EM method for stance identification of topic persons .....	11
3.1 Definition of topic person stance identification .....	11
3.2 A model-based EM method for stance identification of topic persons.....	12
3.2.1 Model-based stance identification of topic persons .....	12
3.2.2 MaxMin initialization algorithm .....	17
3.2.3 Off-topic block elimination .....	18
3.2.4 Weighted correlation coefficient.....	19
3.2.5 Convergence of the EM method.....	20
3.3 Experimental results of EM method.....	21
3.3.1 Data corpus and evaluation metric .....	21
3.3.2 Effect of system components.....	25
3.3.3 Comparison with other clustering methods.....	31
3.3.4 Person stance identification examples.....	35
3.4 Conclusions of the EM method.....	37
4. A topic person stance identification method based on friendship network analysis	38
4.1 Friendship network construction .....	39
4.2 The objective function of SCIFNET .....	44
4.3 Stance expansion .....	44
4.4 Stance refinement .....	49
4.5 Stance-irrelevant topic person detection .....	52
4.6 Experimental results of the SCIFNET.....	54
4.6.1 Dataset .....	54
4.6.2 System component analysis.....	58
4.6.2.1 Friendship orientation threshold.....	58
4.6.2.2 Friendship Orientation Threshold using different perspective.....	63
4.6.2.3 Edge weight evaluation.....	66
4.6.2.4 Stance-oriented correlation coefficient evaluation.....	68
4.6.2.5 The effect of the adoption all the extracted topic persons.....	70
4.6.2.6 The effect of the adoption of the other named entities.....	71
4.6.3 Comparison with other methods.....	72
4.6.3.1 Stance identification evaluation.....	72
4.6.3.2 Stance-irrelevant topic person detection evaluation.....	77
4.6.4 An example of topic person stance identification.....	78
4.7 Conclusions of the SCIFNET .....	80
5 Conclusions .....	81
6. References .....	82



## List of Figures

Figure 1. An example of stance identification of topic persons .....	12
Figure 2. $N \times M$ block-person matrix $BP$ .....	13
Figure 3. The system architecture of model-based EM method.....	13
Figure 4. A document related to the topic “Selection of the new IMF president.”.....	15
Figure 5. The MaxMin initialization algorithm.....	18
Figure 6. The rand index scores under different settings of $\gamma$ .....	26
Figure 7. The rand index with random initialization under $\lambda = 50\%$ .....	27
Figure 8. The rand index with random initialization under $\lambda = 60\%$ .....	27
Figure 9. The rand index with MaxMin under $\lambda = 50\%$ .....	30
Figure 10. The rand index with MaxMin under $\lambda = 60\%$ .....	30
Figure 11. The person stance identification results of the 2008 and 2009 NBA Conference Finals. ....	35
Figure 12. The system architecture.....	39
Figure 13. The stance expansion algorithm.....	45
Figure 14. An example of stance expansion.....	46
Figure 15. An example of stance refinement.....	49
Figure 16. The stance refinement algorithm.....	50
Figure 17. An example of the associations of stance-irrelevant persons.....	54
Figure 18. The effect of parameter $\theta$ under $\lambda = 50\%$ .....	60
Figure 19. The effect of parameter $\theta$ under $\lambda = 60\%$ .....	60
Figure 20. The effect of parameter $\theta$ under $\lambda = 70\%$ .....	61
Figure 21. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 50\%$ .....	62
Figure 22. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 60\%$ .....	62
Figure 23. The F1-score/ratio of the detected stance-irrelevant persons under $\lambda = 70\%$ .....	63
Figure 30. The effect of parameter $\theta$ includes all the extracted person names on Sports topics under $\lambda = 70\%$ .....	70
Figure 31. The F1-score/ratio of the detected stance-irrelevant persons includes all the extracted person names on Sports topics under $\lambda = 70\%$ .....	70
Figure 32. The effect of parameter $\theta$ includes other named entities on Sports topics under $\lambda = 70\%$ .....	71
Figure 33. The F1-score/ratio of the detected stance-irrelevant persons includes other named entities on Sports topics under $\lambda = 70\%$ .....	72
Figure 34. The stance identification result of the 2009 NBA Conference Final ( $\lambda = 70\%$ ) .....	78



## List of Tables

Table 1. The statistics of the evaluation corpus .....	22
Table 2. The stance identification results of the compared methods .....	33
Table 3. The stance identification results for Topic $A_9$ ( $\lambda = 60\%$ , $\beta = 0.4$ with OBE). 36	
Table 4. The stance identification results for Topic $A_{10}$ ( $\lambda = 60\%$ , $\beta = 0.4$ with OBE) 37	
Table 5. The data corpus for the SCIFNET .....	55
Table 6. The lists of <i>Fwords</i> and <i>Owords</i> .....	58
Table 7. Comparison of the edge weighting strategies .....	67
Table 8. Comparison of the correlation coefficient approaches .....	69
Table 9. The rand index performance of the compared methods .....	74
Table 10. The F1 performance of stance-irrelevant topic person detection .....	78



# 1. Introduction

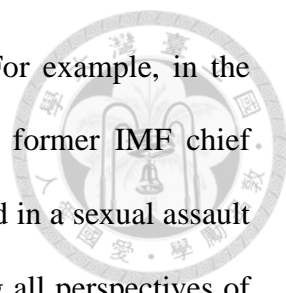
With the prevalence of the Internet and the explosive growth of medium digitization, there are astronomical information on the web. Nowadays, People can easily receive the latest topics, such as global economic trends, politics events, and tournament results, all over the world through the Internet. In general, people are interested in topics that involve competing viewpoints or controversial scenarios. However, they are generally overwhelmed by the huge amount of topic documents which cover every detail of different stances. For example, in the topic of 2011 IMF (International Monetary Fund) presidential selection, Google News<sup>1</sup> collected hundreds of topic documents reporting the development of the campaign. Although the documents reported all perspectives of the topic (i.e., from the interactions between the candidates to the viewpoints of each country's financial representative), readers generally have difficulties assimilating the enormous documents, not to mention understanding different stances of the topic.

To ease the burden of reading a great deal of topic documents, several topic mining techniques have been developed. For instance, Nallapati et al. (2004) grouped topic documents into clusters, each of which presents a theme of a topic. Feng and Allan (2007) extracted informative sentences from themes to summarize a topic. Chen and Chen (2008; 2012) further organized themes and summaries chronologically to depict the storyline of a topic. The techniques successfully condense the content of a topic. However, readers still need to spend a lot of time to digest the generated summaries if they are not familiar with the topic.

Topics basically are associated with persons, times, and places (Nallapati et al., 2004). Identifying the stances of persons in the topics with competing viewpoints (called *topic persons* hereafter) can facilitate readers to construct the background

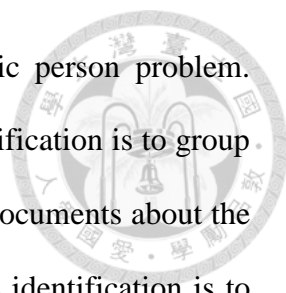
---

<sup>1</sup> <https://news.google.com>

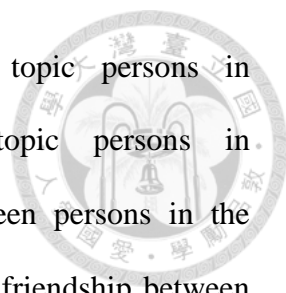


knowledge of the topics and to digest topic documents quickly. For example, in the topic about the selection of the new IMF president in 2011, the former IMF chief Dominique Strauss-Kahn stepped down because he had been charged in a sexual assault case. Google News collected hundreds of topic documents reporting all perspectives of the stances of the following participants: the staff of the Bank of Mexico; the French Minister of Finance; the countries opposed to the French Minister of Finance; and the countries that supported non-European zone candidates. The topic persons with opposing stances competed to have their candidate selected as the new IMF president. If readers knew the persons associated with the four stances, they could have understood the numerous topic documents easily. It is reasonable to ask experts to identify the associations between persons and stances for readers. However, as new topics occur frequently and the corresponding topic documents are posted on the Internet, experts could be overwhelmed by the huge number of documents. The situation would be even worse if the experts were not familiar with the background of a topic. To discover the associations between the persons and stances of an unfamiliar topic, the experts would still need to read the topic documents. That would not be an easy task if there are numerous documents; hence, automatic methods for stance identification of topic persons are essential.

Identifying stances of topic persons is a new research topic. To the best of our knowledge, only Chen et al. (2010; 2012) dealt with the stance identification problem. The authors proposed the use of Principal Component Analysis (PCA) (Barber, 2012) and examine the signs of the entries in the eigenvector associated with the largest eigenvalue to recognize stances of topic persons. The method, however, can simply handle two-stance topics but many topics involve more than two stances in reality.



In this study, we investigate the stance identification of topic person problem. Given a set of topic documents, the task of topic person stance identification is to group topic persons into stance-coherent clusters. For instance, given the documents about the 2011 IMF presidential selection, the task of the topic person stance identification is to identify candidates and their supporters, and other groups of people holding different stances. A challenging issue in the stance identification of topic persons is that stance of the individuals is topic-dependent. For instance, politicians often change their policies for the sake of expediency, so their stances change accordingly. To solve this problem, we propose two unsupervised stance identification approaches. The first approach employs a model-based Expectation-Maximization (EM) method to identify topic persons in an unsupervised manner. As the method only considers the word usage patterns of person names in topic documents, it does not require external knowledge sources and it can capture the feature of person stance's dynamics. A difficulty in EM-based methods is that the results of the methods depend on the initialization of their parameters (Figueiredo & Jain, 2002; Pernkopf & Bouchaffra, 2005). In the study of the first approach, we propose an effective initialization strategy that ensures a stable and accurate stance identification performance. Moreover, we present off-topic block elimination and weighted correlation coefficient techniques to remove the off-topic text blocks and reduce the text sparseness problem respectively. In our model-based EM method, we didn't take the competing semantic into consideration and didn't employ the social network features for identifying topic person stances. We also observed that some of topics contain stance-irrelevant topic persons. Hence, we propose the second approach, namely, a stance community identification based on friendship network (SCIFNET) method to cope with the findings of interest. The SCIFNET constructs a friendship network of topic persons. Nodes in the network represent topic persons.



Edges are established by considering the co-occurrence of topic persons in stance-weighted documents. Then, the co-occurrence of topic persons in stance-weighted documents and the co-neighboring degree between persons in the network are leveraged to define edge weights (i.e., the strength of friendship between persons). An effective community detection algorithm which consists of a stance community expansion algorithm and a stance community refinement technique is presented to group the topic persons into stance-coherent clusters.

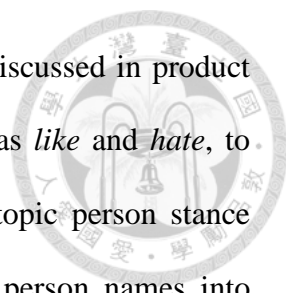
The remainder of this dissertation is organized as follows. We provided the related literatures in Section 2. Then, we describe the methodologies and show experiments in Section 3 and 4. We concluded our findings and future works in section 5.

## **2. Literature review**

In the following, we review research fields related to the topic person stance identification problem.

### **2.1 Opinion mining**

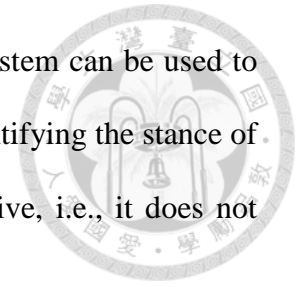
Since our research aims at identifying stances of topic persons, it is related to opinion mining (Liu, 2012), which is also called sentiment analysis. Sentiment analysis usually focuses on discovering textual units with bipolar orientations. However, it differs from sentiment analysis in a number of respects. First, most sentiment analysis approaches identify the polarity of adjectives, adverbs, and verbs because the syntactic constructs generally convey sentimental semantics. For instance, Hatzivassiloglou and McKeown (1997) employed language conjunctions, such as *and*, *or*, and *but*, to judge the polarity of conjoined adjectives. Ganapathibhotla and Liu (2008) investigated the polarity of comparative adjectives (e.g., quick) or adverbs (e.g., quickly) combined with product



features (e.g., run time) to identify the pros and cons of products discussed in product reviews. Ding et al. (2008) also considered sentiment verbs, such as *like* and *hate*, to extract further sentiment comments about a product. In contrast, topic person stance identification considers the stances of topic persons and clusters person names into groups which are nouns that rarely express sentiment information. Second, sentiment analysis generally classifies textual units in terms of a positive or negative orientation, but a person's stance does not have a positive or negative meaning. For example, in political topics, people protest the government or politicians because they made decisions which benefit few of companies or politicians themselves. People may have angry emotion and the consensus of protesting the decision makers. However, the group of the protesters is a stance without negative meaning. Specifically, people with different stances take opposing viewpoint regarding a certain topic, while people in the same stance group reach a consensus or have the same goal. Finally, sentiment analysis usually requires external knowledge sources or human-composed sentiment lexicons. For example, Kim and Hovy (2004), and Hu and Liu (2004) determined a word's polarity by classifying the synonyms and antonyms of the word in WordNet (Miller et al., 1990); while Ku et al. (2006) dealt with Chinese sentiment analysis by considering the sentiment words in the General Inquirer lexicon (Stone et al., 1966). However, no external knowledge source is available for topic person stance identification research because a person's stance is dynamic and topic-dependent. The property of topic-dependence and the lack of knowledge sources make the topic person stance identification task a challenging research issue.

In addition, Godbole et al. (2007) developed a system to extract the positive or negative comments about a person from weblogs and news articles. The authors manually compiled a list of sentiment words and then extended the list with WordNet

(Miller et al., 1990) to calculate the person's polarity score. The system can be used to measure a person's reputation. By contrast, we focus mainly on identifying the stance of a group of topic persons. The stance is neither positive nor negative, i.e., it does not have a positive or negative meaning.

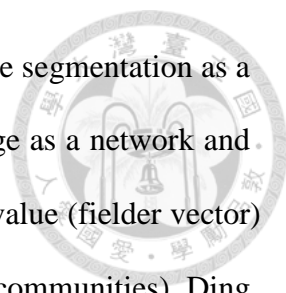


## **2.2 Community detection**

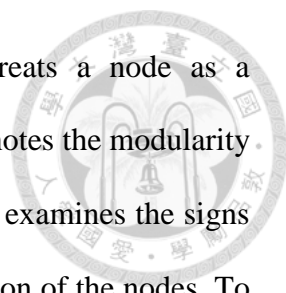
As the person stance identification is to cluster persons into stance-coherent clusters. Our research is also related to community detection (Papadopoulos et al., 2012). Specifically, given a network of interests, the task of community detection is to identify sub-networks so that each of which represents a coherent community (Clauset et al., 2004; Girvan & Newman, 2002; Newman, 2001, 2004; Newman & Girvan, 2004). For instance, given a social network, community detection identifies groups of people with similar preferences (Papadopoulos et al., 2012). Basically, community detection methods partition the given network into sub-networks (i.e., communities) in accordance with the principle that maximizes the association within each sub-network, while minimizing the association between them (Shi & Malik, 2000). In the following sub-sections, we review two main community detection approaches, namely, the eigen-based community detection approach and the iterative clustering approach.

### **2.2.1 Eigen-based community detection approach**

One family of the eigen-based community detection approach is spectral clustering which makes use of the eigenvectors of the Laplacian matrix (Donath & Hoffman, 1973) to find appropriate partitions of a network. Given a network, the Laplacian matrix is derived by subtracting the adjacency matrix  $A$  from the diagonal matrix  $D$ . The entry  $a_{i,j}$  in  $A$  is 1 if node  $i$  and node  $j$  are connected, otherwise it is 0, and the entry  $d_{i,i}$  in  $D$  is the



degree of node  $i$  in the network. Shi and Malik (2000) modeled image segmentation as a community detection problem. The authors first represented an image as a network and employed the eigenvector associated with the second smallest eigenvalue (fielder vector) of the Laplacian matrix to identify significant image segments (i.e., communities). Ding et al. (2001) employed spectral clustering to cluster a set of documents. The authors constructed a word-document matrix  $X$  in which entries are the mutual information (Manning et al., 2008) between words and documents. Then, a document network is constructed by considering each document as a node. The connection between nodes is represented by the weighted matrix  $W=X^T X$ . The network is partitioned by using the fielder vector of the matrix  $W$ . The authors also introduced the Mcut metric to evaluate the partitioned network. The metric is integrated with a linkage-based refinement technique to improve the quality of the network partition. A limitation of the above methods is that they generally make balanced cuts in partitioning a network, that is, the detected communities in the network need to be with a similar size. In practice, however, communities are with different sizes and magnitudes so that the balanced cut requirement is irrational (Newman, 2006; White & Smyth, 2005). To relax this limitation, White and Smyth (2005) developed a spectral clustering algorithm which maximizes the modularity (Newman & Girvan, 2004) of a network partition. The larger the value is, the better the quality of the network partition will be. The authors formulated the modularity maximization problem as a quadratic assignment problem and solved it analytically using an eigen-decomposition method. Specifically, the method constructs an eigenvector matrix  $U_K$  where the columns are the eigenvectors of the matrix  $L_Q$  derived from the modularity maximization problem. Then, the row vectors of  $U_K$  are clustered by using the k-means algorithm (Manning et al., 2008) to find an appropriate network partition. Newman (2006) developed an efficient algorithm to

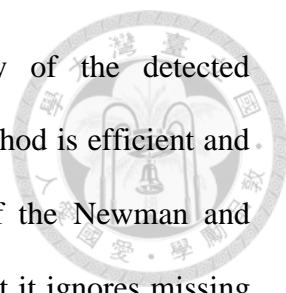


detect communities within a network. Initially, the algorithm treats a node as a community and constructs a modularity matrix  $B$  where entry  $B_{i,j}$  denotes the modularity between the community  $i$  and the community  $j$ . Then, the algorithm examines the signs of the entries in the principal eigenvector of  $B$  to identify the affiliation of the nodes. To polish detected communities, i.e., subgraphs in the network, the algorithm further examines the modularity changed by moving nodes between subgraphs and moves all the nodes that increase the modularity. Anchuri and Magdon-Ismai (2012) investigated signed networks in which nodes are connected by positive or negative edges. They modified the modularity to incorporate negative edges into it and constructed a modularity matrix for a signed network. Communities are detected by examining the signs in the matrix's eigenvector associated with the largest eigenvalue. In addition, a refinement method based on the modified modularity is developed to calibrate the membership of the nodes.

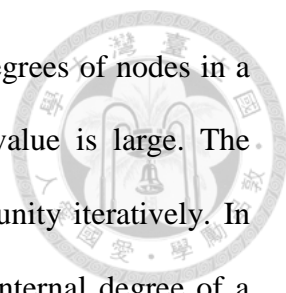
### **2.2.2 Iterative clustering approach**

Another popular approach of community detection is iterative clustering. Girvan and Newman (2002) devised a hierarchical clustering algorithm which measures the betweenness of edges for community detection. The betweenness of an edge denotes the number of shortest paths between pairs of nodes that run through the edge. The algorithm iteratively decomposes the network by removing the edge with the highest betweenness until a specific number of communities have been detected. Newman and Girvan (2004) proposed a betweenness-based method for community detection. The authors also developed a measurement called modularity to evaluate the quality of the detected communities. Meanwhile, Newman (2004) proposed a modularity-based community detection algorithm which initializes each node as a community. Then, the





algorithm iteratively merges communities until the modularity of the detected communities reaches a local optimal. The authors show that the method is efficient and the quality of the detected communities is comparable to that of the Newman and Girvan's method. A problem of the modularity-based method is that it ignores missing edges within a community. In other words, the modularity only measures how good the discovered community structure fits the existing edges (Chen et al., 2009a). In reality, it is difficult to acquire all information about the analyzed network. So, the network may miss informative edges that deteriorate community detection performance. To resolve the problem, Chen et al. (2009a) developed a new measurement, called Max-Min modularity which considers missing edges to improve the quality of community detection. Xu et al. (2007) also proposed an iterative clustering algorithm for community detection. For every node pair, the algorithm first computes the ratio of co-neighbors between them. A node is considered the core of a community if the number of the high co-neighbor ratios between it and other nodes are also high. The algorithm expands communities from core nodes and iteratively labels their neighbors the same communities. It is worthy to note that the algorithm can identify the hub nodes which function as a bridge to connect to different communities. In social networks, the hub nodes may play an important role in viral marketing. Yang et al. (2007) developed an iterative bipartition method called FEC (Finding and Extracting a Community) for detecting communities in a signed network. The method first conducts a random walk on the network to measure the probability of reaching a node. Afterward, an adjacency matrix is constructed by sorting the nodes in accordance with their reaching probabilities. The algorithm then iteratively identifies a cutting point in the matrix to bipartition the network such that the positive edges within the partitioned sub-networks and the negative edges between the sub-networks are dense. Chen et al. (2009b)



developed the metric  $L$  which leverages the internal and external degrees of nodes in a community. A detected community is considered good if its  $L$  value is large. The authors also developed a two-phrase algorithm to expand a community iteratively. In phrase one, the nodes whose degrees are larger than the average internal degree of a community are identified. In the second phrase, the nodes are examined and included in the community if their inclusions increase the community's  $L$  value. Their experiments showed that the communities detected by using  $L$  are superior. Traag and Bruggeman (2009) modified the modularity to incorporate negative edges of a network. The modularity is incorporated the Potts model (Wu, 1982) to detect communities. Yang et al. (2009) integrated link structure with content analysis for community detection. They presented a popularity-based link model to measure the strength between nodes and employed an EM process to learn the memberships of nodes. Gao et al. (2010) developed a generative model, called CODA (Community Outlier Detection Algorithm), to detect communities and outliers. The model employs the hidden Markov random fields (Barber, 2012) to compute the importance of network structure. Moreover, the algorithm sorts nodes in terms of objective values to identify outliers.

Technically, our second approach, SCIFNET, differs from existing community detection in many respects. First, community detection generally partitions the entire network. In the topic person stance identification task, however, stance-irrelevant persons (i.e., the persons with no stance) exist and they do not belong to any community. Our SCIFNET can detect the stance-irrelevant persons. Second, the networks analyzed by community detection approaches are generally pre-defined. In the SCIFNET, the friendship network of topic persons is derived automatically from topic documents. Third, the above studies usually only consider the link structures of the nodes but ignore other features of the nodes (e.g., the co-occurrence patterns of the nodes in documents).

We not only consider the link structure but also consider the co-occurrence patterns of the persons in our SCIFNET. Finally, the networks of community detection do not convey competing semantics. By contrast, the SCIFNET considers friendship orientations, and identifies friendly and opposing relationships between topic persons.

### **3. A model-based EM method for stance identification of topic persons**

In this section, we define the problem of topic person stance identification, and then introduce our model-based EM methods for identifying the stances of topic persons.

#### **3.1 Definition of topic person stance identification**

Given a set of documents about a topic that involves competing viewpoints with  $K$  stances, the task of stance identification of topic persons involves clustering the persons mentioned in the documents into  $K$  stance-coherent groups. For example, Figure 1 shows documents related to the selection of the new IMF president in 2011. The stance identification method clusters the mentioned persons into four stance-coherent groups: the staff of the Bank of Mexico, the French Minister of Finance and her representatives, the South African delegates opposed to the French Minister of Finance, and the country delegates that supported non-European zone candidates. We posit that identifying stance-coherent groups of topic persons can help readers construct the knowledge background of a topic and help them comprehend the topic documents quickly. In the following subsections, we detail our proposed methods.

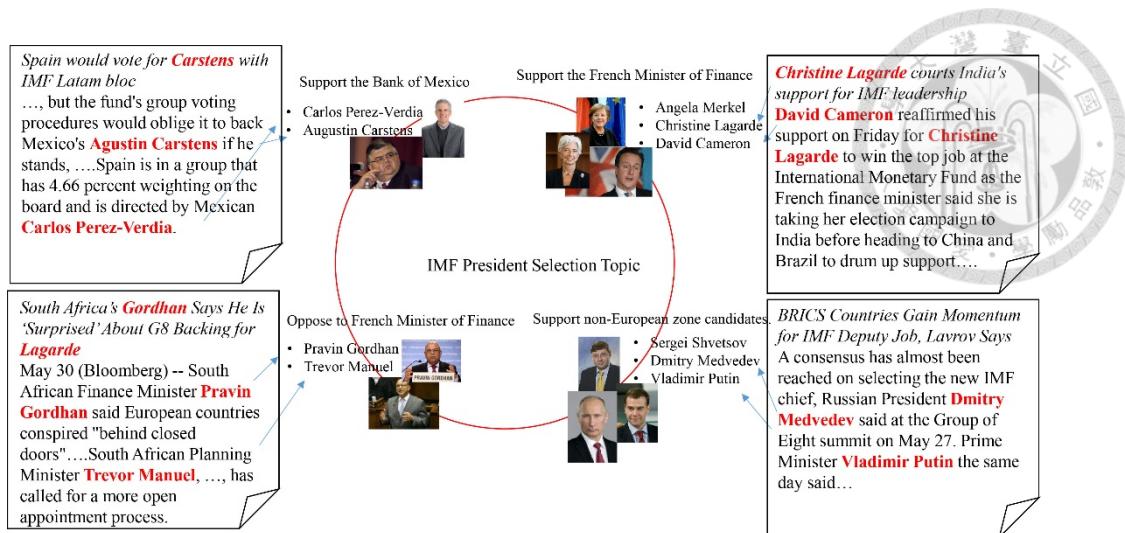


Figure 1. An example of stance identification of topic persons

## 3.2 A model-based EM method for stance identification of topic persons

### 3.2.1 Model-based stance identification of topic persons

To identify the stances of topic persons, we first decompose the documents into a set of non-overlapping blocks  $B = \{b_1, \dots, b_N\}$ . A block is a content coherent unit, i.e., a document or a paragraph. Let  $P = \{p_1, \dots, p_M\}$  represent a set of person names mentioned in  $B$  (i.e., topic persons). Then, the topic can be described by an  $N \times M$  block-person association matrix  $BP$ , as shown in Figure 2. The  $j$ -th row in  $BP$  represents a block  $b_j$ . It is an  $M$ -dimensional vector whose  $i$ 'th entry, denoted by  $b_{j,i}$ , is the frequency of person name  $p_i$  in block  $b_j$ . Meanwhile, a topic person  $p_i$  is represented as a column in  $BP$ . The column is an  $N$ -dimensional frequency vector whose  $j$ 'th entry, denoted by  $p_{i,j}$ , is the frequency of person name  $p_i$  in block  $b_j$ . Figure 3 shows the system architecture below. First of all, we introduce the EM step and detail off-topic block elimination and initialization algorithm later.

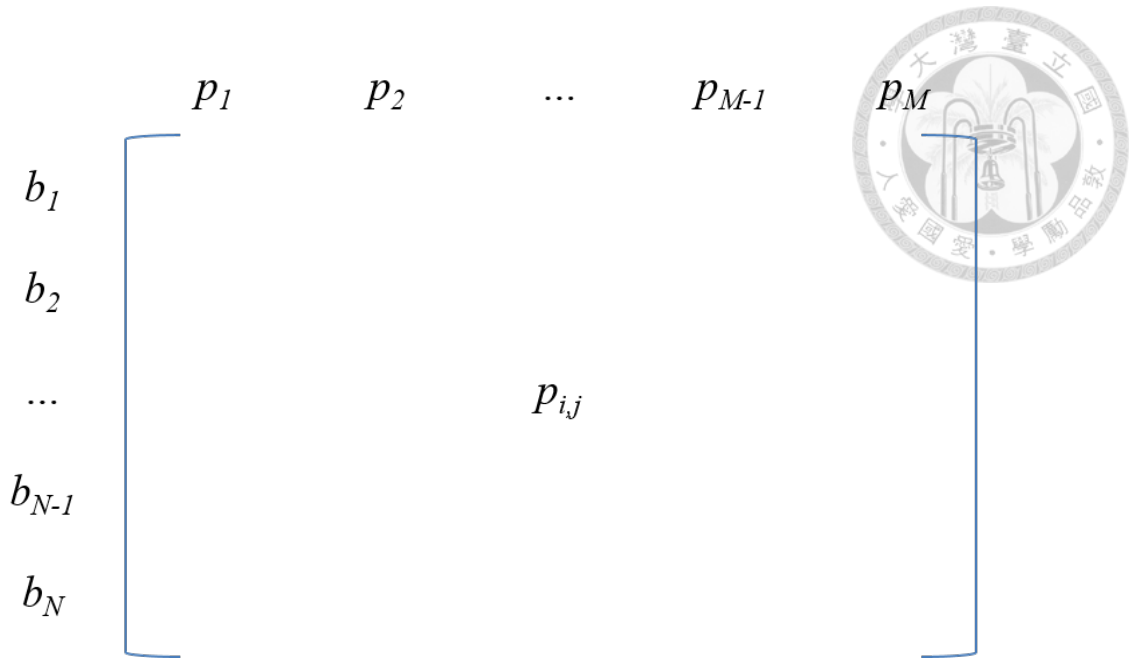


Figure 2.  $N \times M$  block-person matrix  $BP$

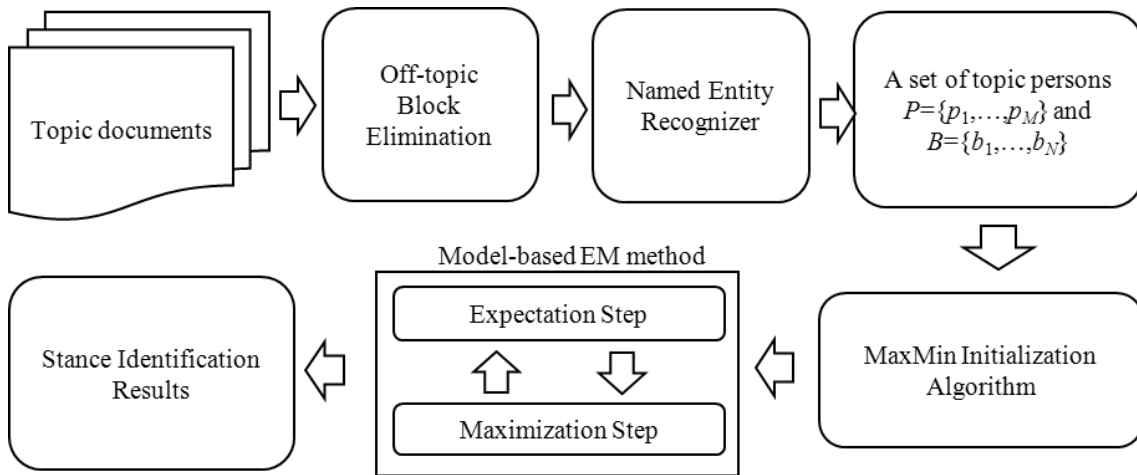


Figure 3. The system architecture of model-based EM method

After modeling the topic persons as high-dimensional frequency vectors, we utilize a model-based EM method to identify their stances. Let  $\theta = \{(\alpha_1, \omega_1), \dots, (\alpha_K, \omega_K)\}$  represent the stance model, where  $\alpha_k$  is stance-group  $k$ 's weight, such that  $\sum \alpha_k = 1$ . Here,  $\omega_k$  is an  $N$ -dimensional representative vector of stance-group  $k$ . It is a weighted centroid of the stance-group members' frequency vectors. Therefore, the  $l$ 'th entry, denoted by  $\omega_{k,l}$ , is the weight of the block  $b_l$  of stance-group  $k$ . We formulate the stance identification of topic person problem as follows:

$$\hat{\theta} = \arg \max_{\theta \in \text{search space}} P(\theta | P). \quad (1)$$

Then, based on Bayes' theorem, we expand the above equation to the following form:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta \in \text{search space}} \frac{P(\theta)P(P | \theta)}{P(P)} \\ &= \arg \max_{\theta \in \text{search space}} P(\theta)P(P | \theta). \end{aligned} \quad (2)$$

As the number of stance models is infinite, it is reasonable to assume that all models have the same prior probability  $P(\theta)$  (Mitchell, 1997). Hence, Eq. (2) can be rewritten as follows:

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta \in \text{search space}} P(P | \theta) \\ &= \arg \max_{\theta \in \text{search space}} \prod_{i=1}^M \sum_{k=1}^K \alpha_k P(p_i | \omega_k). \end{aligned} \quad (3)$$

In other words, the stance model that is searched should contain the maximum likelihood of the person name occurrences. To search for the stance model, we need to define  $P(p_i | \omega_k)$ . Most topic documents focus on individual stances because the documents are published in chronological order (Nallapati et al., 2004). When an event occurs, the topic documents usually focus on the first stance. Subsequently, other stances will be reported in different topic documents to show the development of the topic. The chronological property corresponds with the findings of Kanayama and Nasukawa (2006) who validated that text units with the same polarity tend to occur (not occur) jointly to make contexts coherent. Consequently, persons mentioned in the same document are likely to be associated with the same stance. Moreover, if the occurrences of a person name are coincident with those of a stance-coherent group, the person can be regarded as a member of that group and therefore has a high  $P(p_i | \omega_k)$ .

Germany's **Merkel: Lagarde** Is Ideal Embodiment Of Economic, Political Experience For IMF  
 (<http://english.capital.gr/news.asp?id=1208771>)

SINGAPORE (Dow Jones)--German Chancellor **Angela Merkel** Thursday said French Finance Minister **Christine Lagarde** has the economic and political experience to head the International Monetary Fund. "If I look at the personality of **Christine Lagarde**, as a finance minister, she enjoys an excellent reputation worldwide, and in many ways is an ideal embodiment of economic and political experience," **Merkel** said. **Merkel** said that in the long-run, it is unacceptable to think that a European IMF head and a U.S. World Bank head would be automatic, but now might not be the right time to alter that model. "Since the IMF is very deeply involved in the euro matters, there could be good reasons for not saying right away that a European candidate is out of the question," the German chancellor said. **Merkel** said she hoped emerging countries would take "an objective and unbiased look at" **Lagarde** for the post. **Merkel's** remarks came as she delivered a lecture hosted by the Institute of Southeast Asian Studies in Singapore. By tradition a European has always headed the IMF, but many in Asia and other emerging economies say the practice is outdated. The IMF job came open after the resignation of **Dominique Strauss-Kahn**, a former French finance minister arrested last month on charges he attempted to rape a maid in a New York hotel. **Strauss-Kahn** denies all charges.

Figure 4. A document related to the topic "Selection of the new IMF president."

For example, in Figure 4, the document entitled "Germany's Merkel: Lagarde Is Ideal Embodiment of Economic, Political Experience for IMF" reports an important event where Angela Merkel declared her support for Christine Lagarde as the new IMF president. In the document, Merkel and Lagarde are mentioned frequently to explain the important event. We learn  $P(p_i|\omega_k)$  from topic blocks and use the following correlation coefficient to discover the joint behavior of topic persons.

$$corr(p_i, \omega_k) = \frac{\sum_{l=1}^N (p_{i,l} - \tilde{p}_i) * (\omega_{k,l} - \tilde{\omega}_k)}{\sqrt{\sum_{l=1}^N (p_{i,l} - \tilde{p}_i)^2} \sqrt{\sum_{l=1}^N (\omega_{k,l} - \tilde{\omega}_k)^2}}, \quad (4)$$

where  $corr(\cdot)$  denotes the correlation coefficient between the representative vectors of topic person  $p_i$  and stance-group  $k$ ; and  $\tilde{p}_i = (\sum_{l=1}^N p_{i,l})/N$ ; and  $\tilde{\omega}_k = (\sum_{l=1}^N \omega_{k,l})/N$  are the average frequencies of topic person  $p_i$  and the members of stance-group  $k$  respectively. The range of the correlation coefficient is within  $[-1,1]$ . It represents the degree of joint behavior of topic person  $p_i$  and stance-group  $k$  under the decomposed blocks. A positive value means that the  $p_i$  and the members of stance-group  $k$  tend to occur (not occur) jointly in the topic blocks. Conversely, a negative value indicates that the occurrences of

the  $p_i$  and the members of stance-group  $k$  are negatively correlated. To avoid negative probabilities in  $P(p_i|\omega_k)$ , we define the following function to convert the range of the correlation coefficient:

$$strength(p_i, \omega_k) = \frac{(corr(p_i, \omega_k) + 1)}{2}. \quad (5)$$

The range of the conversion function  $strength(.)$  is within  $[0,1]$ . The function returns 1 when the topic person  $p_i$  and the members of stance-group  $k$  are positively correlated, and 0 when they are negatively correlated. We define  $P(p_i|\omega_k)$  as follows:

$$P(p_i | \omega_k) = \frac{strength(p_i, \omega_k)}{\sum_{j=1}^M strength(p_j, \omega_k)}. \quad (6)$$

The denominator in Eq. (6) is a normalization factor; hence, topic persons positively correlated with stance-group  $k$  would belong to the stance whose centroid is  $\omega_k$ . Then, our objective (as defined in Eq. (3)) is to cluster topic persons into positively correlated groups.

Let  $\langle h_{i,1}, \dots, h_{i,K} \rangle$  denote a sequence of binary variables of topic person  $p_i$ . Here,  $h_{i,k} = 1$  if person  $p_i$  belongs to stance-group  $k$ ; otherwise,  $h_{i,k} = 0$ . As stance identification of topic persons is an unsupervised problem, the values of the variables are unobserved. We exploit an EM method to search for appropriate person stances. First, we randomly initialize the model parameters, and then execute the following EM steps iteratively until convergence.

$$E - step : E[h_{i,k}] = \frac{\alpha_k * P(p_i | \omega_k)}{\sum_{j=1}^K \alpha_j * P(p_i | \omega_j)}. \quad (7)$$

$$M - step : \alpha_k = \frac{\sum_{i=1}^M E[h_{i,k}]}{M} \quad \text{and} \quad \omega_k = \frac{\sum_{i=1}^M E[h_{i,k}] * p_i}{\sum_{i=1}^M E[h_{i,k}]}. \quad (8)$$

The E-step uses the current stance model to compute the expectation of an



unobserved variable  $h_{i,k}$ . Then, the  $M$ -step re-computes the stance model as the maximum likelihood estimates given all the calculated expectations. When convergence occurs,  $E[h_{i,k}]$  indicates the probability that topic person  $p_i$  belongs to stance-group  $k$ . We then assign topic person  $p_i$  to the stance with the maximum probability.

### 3.2.2 MaxMin initialization algorithm

A shortcoming of model-based EM methods is that the result depends on the model's initialization (Figueiredo & Jain, 2002; Pernkopf & Bouchaffra, 2005). As mentioned earlier, the proposed stance identification method utilizes a random stance model and iterates the EM operations to improve the stance identification result. Here, we present an effective model initialization that yields stable and accurate stance identification results.

The initialization algorithm selects representative topic persons of  $K$  stances and uses their frequency vectors to initialize  $\omega_k$ , as shown in Figure 5. Let  $I$  denote the set of selected persons. Initially, the set is empty. The algorithm first selects the person who has the maximum correlation with the topic persons. That person is regarded as the most representative topic person, so he/she is added to  $I$ . The correlation between the persons in  $I$  should be low to distinguish between different stances; hence, the algorithm iteratively selects  $K-1$  persons that have the minimum correlation with the persons already in  $I$ . As the algorithm first selects the person with the maximum correlation and eventually selects the person name with the minimum correlation, we call it the MaxMin initialization algorithm. After selecting  $K$  persons, we take their frequency vectors as the initial  $\omega_k$ 's and initiate the EM procedure.



**MaxMin initialization algorithm :**

$$I = \phi$$

$$p = \arg \max_{p_i} \sum_{j=1, j \neq i}^M \text{corr}(p_i, p_j)$$

$$I = I \cup \{p\}$$

for  $k = 2$  to  $K$

$$p = \arg \min_{p_i, p_i \notin I} \sum_{p_j \in I} \text{corr}(p_i, p_j)$$

$$I = I \cup \{p\}$$

end for

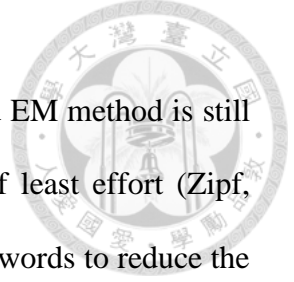
**Initialize**  $\omega_k$ 's with the representative vectors of  $I$ .

$$\alpha_k = \frac{1}{k}$$

Figure 5. The MaxMin initialization algorithm

### 3.2.3 Off-topic block elimination

While collecting the experimental data, we observed that topic blocks are sometimes off-topic. Since topic person names tend to be jointly absent from off-topic blocks, including the blocks in the EM operations would cause the EM method to overestimate the correlation between opposing persons and stances. Therefore, to reduce the influence of off-topic blocks, we implement an off-topic block elimination (OBE) procedure. For each topic, we construct a topic representative vector  $\underline{B}$  by averaging all blocks  $b_l$ . The  $i$ 'th entry of the topic representative vector, denoted by  $B_i$ , is the average frequency of person name  $p_i$  in all the blocks. Then, we use the cosine function (Manning & Schütze, 1999) to calculate the similarity between a topic block and the topic representative vector. Blocks whose cosine similarity to the representative vector  $\underline{B}$  is lower than a predefined threshold  $\gamma$  are deemed off-topic blocks and excluded from the EM procedure.

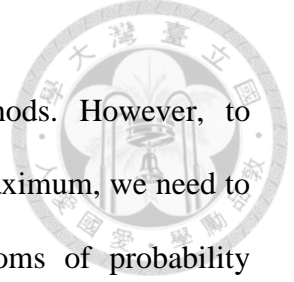


### 3.2.4 Weighted correlation coefficient

Although OBE reduces the number of off-topic blocks, the proposed EM method is still affected by the text sparseness problem. Based on the principle of least effort (Zipf, 1949), document authors tend to use a small vocabulary of common words to reduce the reading (resp. writing) effort that readers (resp. authors) must expend. Consequently, the frequency distribution of person names follows Zipf's law (Zipf, 1949); that is, there are only a few frequent (important) persons, and most person names rarely occur in topic blocks. Hence, many frequency vectors of topic persons contain a lot of zeros, which affect the calculation of the correlation coefficient. The absence of person names from topic blocks could cause overestimation (or underestimation) of the correlation between persons and stances. To address the problem, we propose the following weighted correlation coefficient, called  $wcorr(\cdot)$ , to weight absent blocks:

$$wcorr(p_i, \omega_k) = \frac{\left( (1-\beta) \sum_{b \in co(i,k)} (p_{i,b} - \tilde{p}_i) * (\omega_{k,b} - \tilde{\omega}_k) + \beta \sum_{b \in B-co(i,k)} (p_{i,b} - \tilde{p}_i) * (\omega_{k,b} - \tilde{\omega}_k) \right)}{\sqrt{(1-\beta) \sum_{b \in co(i,k)} (p_{i,b} - \tilde{p}_i)^2 + \beta \sum_{b \in B-co(i,k)} (p_{i,b} - \tilde{p}_i)^2} * \sqrt{(1-\beta) \sum_{b \in co(i,k)} (\omega_{k,b} - \tilde{\omega}_k)^2 + \beta \sum_{b \in B-co(i,k)} (\omega_{k,b} - \tilde{\omega}_k)^2}}, \quad (9)$$

where  $co(i,k)$  denotes the set of blocks whose frequencies in both  $p_i$  and  $\omega_k$  are non-zero. In other words, if we treat  $\omega_k$  as the representative person of stance  $k$ ,  $co(i,k)$  denotes the set of blocks in which person  $p_i$  and stance  $k$  co-occur. Parameter  $\beta$ , whose range is within  $[0,1]$ , weights the influence of non-co-occurring blocks when we calculate the correlation coefficient. A large  $\beta$  value means a non-co-occurring block is important for stance identification. When  $\beta = 0.5$ , the equation is equivalent to the standard correlation coefficient. Like the correlation coefficient, the range of  $wcorr(\cdot)$  is within  $[-1,1]$ . We therefore apply Eq. (5) to avoid negative probabilities when calculating  $P(p_i|\omega_k)$ . In the experiments, we will examine the effect of the value of  $\beta$  on stance identification of topic persons.



### 3.2.5 Convergence of the EM method

Wu (1983) proved the convergence of model-based EM methods. However, to guarantee that the iterative EM steps of our method reach a local maximum, we need to prove that the defined  $P(p_i|\omega_k)$  (i.e., Eq. (6)) satisfies the axioms of probability (Bartoszynski & Niewiadomska-Bugaj, 1996).

*Axiom 1:  $P(p_i|\omega_k)$  is non-negative.*

Proof: As the range of  $strength(\cdot)$  is within  $[0,1]$ , the numerator in Eq. (6) can never be negative. However, to show that  $0 \leq P(p_i|\omega_k)$ , we need to prove that the denominator in the equation is always greater than zero. The (weighted) correlation coefficient is identical to the (weighted) inner product when the frequency vectors are mean-normalized unit vectors (Chen et al., 2010; 2012). Therefore, we convert the denominator in Eq. (6) as follows:

$$\begin{aligned}
 & \sum_{i=1}^M strength(p_i, \omega_k) \\
 &= \sum_{i=1}^M \left[ \frac{corr(p_i, \omega_k) + 1}{2} \right] \\
 &= \frac{1}{2} \sum_{i=1}^M \left[ \left( \frac{1}{\sum_{l=1}^M E[h_{l,k}]} \sum_{l=1}^M E[h_{l,k}] \underline{p}_i \cdot \underline{p}_l \right) + 1 \right],
 \end{aligned} \tag{10}$$

where  $\underline{p}_i$  and  $\underline{p}_l$  represent the mean-normalized unit vectors of  $p_i$  and  $p_l$ , respectively. Equation (10) only reaches its minimum, when all the inner products between  $\underline{p}_i$  and  $\underline{p}_l$  are  $-1$ . However, as  $\underline{p}_l$  represents the mean-normalized unit vector of a topic person, there must be an inner product whose value is 1 (i.e.,  $\underline{p}_i \cdot \underline{p}_l = 1$  when  $p_i = p_l$ ). As a result, the denominator in Eq. (6) is always greater than 0, so  $P(p_i|\omega_k)$  must be non-negative.

*Axiom 2: The sum of all possible  $P(p_i|\omega_k)$  is 1.*



Proof: As the sample space of  $P(p_i|\omega_k)$  is  $P$ , the sum of all possible  $P(p_i|\omega_k)$  is calculated as follows:

$$\begin{aligned}
 \sum_{i=1}^M P(p_i | \omega_k) &= \sum_{i=1}^M \frac{\text{strength}(p_i, \omega_k)}{\sum_{l=1}^M \text{strength}(p_l, \omega_k)} \\
 &= \frac{\sum_{i=1}^M \text{strength}(p_i, \omega_k)}{\sum_{l=1}^M \text{strength}(p_l, \omega_k)} \\
 &= 1.
 \end{aligned} \tag{11}$$

*Axiom 3: For a sequence of mutually exclusive events, the probability of at least one of the events occurring is simply the sum of their respective probabilities.*

Proof: Let each sample point  $p_i$  of  $P(p_i|\omega_k)$  represents an event. As  $p_i$  are individual sample points in  $P$ ,  $p_i \cap p_j$  is empty and the events are mutually exclusive. Moreover, as a topic consists of  $K$  stances, the probability that at least one person will be generated by  $\omega_k$  is 1, i.e.,  $P(p_1 \cup \dots \cup p_M | \omega_k) = 1 = \sum_{i=1}^M P(p_i|\omega_k)$ . Thus,  $P(p_i|\omega_k)$  satisfies Axiom 3.

Because the defined  $P(p_i|\omega_k)$  satisfies the axioms of probability, the iterative EM steps must converge to an appropriate stance model (Wu, 1983).

### 3.3 Experimental results of EM method

In the subsection, we introduce the data corpus and the evaluation metric used in the experiments; assess the effects of OBE, the weighted correlation coefficient, and the MaxMin initialization algorithm; compare the model-based EM method's performance with those of well-known clustering algorithms; and show examples of stance identification.

#### 3.3.1 Data corpus and evaluation metric

In text mining, evaluations are normally based on official benchmarks. However, to the best of our knowledge, there are no official corpora for identifying the stances of topic

persons because the research field is relatively new. We therefore compiled a data corpus to evaluate our method. As shown in Table 1, the corpus comprises sixteen topics with different stances. All the topic documents were downloaded from Google News. We selected the topics for evaluation because they are all related to global news events, so readers can comprehend the stance identification examples presented in Section 3.3.4 without specific cultural or background knowledge. To compare our method with Chen et al.'s bipolarization approach (2010; 2012), we prepared eight topics with two stances (i.e., Topics  $A_1 \sim A_8$ ). Topics  $A_1$  to  $A_4$ , which are related to business topics, comprise 123, 74, 154, and 48 news documents respectively. Topics  $A_5 \sim A_8$  are related to four sports tournaments. We also collected topics for four stances. Topics  $A_9 \sim A_{12}$  are related to the NBA Conference Finals from 2008 to 2011. Each topic involves four basketball teams that competed for the title. Topics  $A_{13} \sim A_{16}$  are global business topics.

Table 1. The statistics of the evaluation corpus

ID	Date	Topic Description				
		# of topic documents	# of extracted persons	# of evaluated persons ( $\lambda = 50\%$ )	# of evaluated persons ( $\lambda = 60\%$ )	
Stance Description						
$A_1$	2010/7/18-2010/7/22	Smartphone manufacturers deny Apple's reception claim	123	74	3	5
		<ul style="list-style-type: none"> <li>● Support Apple's reception claim</li> <li>● Deny Apple's reception claim</li> </ul>				
$A_2$	2010/8/4-2010/8/6	Google-Verizon deny tiered-web deal report	74	53	5	7
		<ul style="list-style-type: none"> <li>● Oppose the cooperation of Google and Verizon</li> <li>● Support the cooperation of Google and Verizon</li> </ul>				
$A_3$	2010/6/1-2010/6/3	Prudential's shareholders oppose buying AIG's Asian Unit	154	93	2	3
		<ul style="list-style-type: none"> <li>● Support buying AIG's Asian Unit</li> <li>● Oppose buying AIG's Asian Unit</li> </ul>				
$A_4$	2010/1/13-2010/1/15	Google ends four years of censoring the Web for China.	48	103	9	13
		<ul style="list-style-type: none"> <li>● Support Google's decision to quit the China market</li> <li>● Support China's censorship of Google content</li> </ul>				
$A_5$	2009/6/4-2009/6/16	The 2009 NBA Finals				

		411	423	6	8
		<ul style="list-style-type: none"> <li>● Lakers basketball team competing in the 2009 NBA championship</li> <li>● Magic basketball team competing in the 2009 NBA championship</li> </ul>			
A <sub>6</sub>	2010/4/1-2010/4/5	The opening game of the 2010 MLB season			
		33	77	5	7
		<ul style="list-style-type: none"> <li>● Washington Nationals team</li> <li>● Philadelphia Phillies team</li> </ul>			
A <sub>7</sub>	2010/6/4-2010/6/19	The 2010 NBA Finals			
		87	141	5	8
		<ul style="list-style-type: none"> <li>● Lakers basketball team</li> <li>● Celtics basketball team</li> </ul>			
A <sub>8</sub>	2010/7/10-2010/7/12	The 2010 World Cup Final			
		166	214	9	12
		<ul style="list-style-type: none"> <li>● Dutch team competing in World Cup Championship</li> <li>● Spanish team competing in World Cup Championship</li> </ul>			
A <sub>9</sub>	2008/5/20-2008/5/30	The 2008 NBA Conference Finals			
		119	77	8	12
		<ul style="list-style-type: none"> <li>● Celtics basketball team</li> <li>● Pistons basketball team</li> <li>● Lakers basketball team</li> <li>● Spurs basketball team</li> </ul>			
A <sub>10</sub>	2009/5/19-2009/5/30	The 2009 NBA Conference Finals			
		78	147	11	17
		<ul style="list-style-type: none"> <li>● Cavaliers basketball team</li> <li>● Magic basketball team</li> <li>● Lakers basketball team</li> <li>● Nuggets basketball team</li> </ul>			
A <sub>11</sub>	2010/5/16-2010/5/30	The 2010 NBA Conference Finals			
		166	162	12	17
		<ul style="list-style-type: none"> <li>● Celtics basketball team</li> <li>● Magic basketball team</li> <li>● Suns basketball team</li> <li>● Lakers basketball team</li> </ul>			
A <sub>12</sub>	2011/5/14-2011/5/27	The 2011 NBA Conference Finals			
		292	135	9	13
		<ul style="list-style-type: none"> <li>● Bulls basketball team</li> <li>● Heat basketball team</li> <li>● Mavs basketball team</li> <li>● Thunder basketball team</li> </ul>			
A <sub>13</sub>	2011/5/27-2011/6/5	IMF meeting to select a new president			
		150	66	5	11
		<ul style="list-style-type: none"> <li>● Support Agustín Carstens's selection as president of the IMF</li> <li>● Support Christine Lagarde's selection as president of the IMF</li> <li>● Oppose Christine Lagarde's selection as president of the IMF</li> <li>● Support the selection of a non-European zone candidate as president of the IMF</li> </ul>			
A <sub>14</sub>	2011/6/6-2011/6/10	2011 OPEC meeting to set oil production quotas			
		118	167	22	31
		<ul style="list-style-type: none"> <li>● Support an increase in oil production quotas</li> <li>● Oppose an increase in oil production quotas</li> <li>● Neutral on the topic (e.g., OPEC officials)</li> <li>● Analysts providing objective analyses</li> </ul>			
A <sub>15</sub>	2011/6/6-2011/6/11	Greek Financial Crisis			

		135	116	12	19
		<ul style="list-style-type: none"> <li>● The countries involved in the debt crisis</li> <li>● People that provided advice to help the above countries restructure their economies</li> <li>● Support opinion of the European Central Bank (ECB)</li> <li>● Support Germany, which disagreed with the ECB's opinions</li> </ul>			
		Microsoft and i4i lawsuit over patent violation			
		92	32	9	15
A <sub>16</sub>	2011/6/9-2011/6/16	<ul style="list-style-type: none"> <li>● Support Microsoft</li> <li>● Support Canadian software company i4i</li> <li>● The judges who decided the outcome of the lawsuit</li> <li>● Companies that tried to take advantage of Microsoft and i4i</li> </ul>			

For each of the sixteen topics, we used the Stanford Named Entity Recognizer<sup>2</sup> to extract all the person names mentioned in the topic documents. Given an input text, the recognizer extracts all possible named entities from the text and tags each one with a person name, a location name, or an organization name. We used the extracted person names for evaluation. As there is no perfect named entity recognition approach, the recognizer identified false person name entities, such as misspelled names. To evaluate the performance of stance identification of topic persons, we removed false entities comprised of the names of persons and the names of organizations (or locations) because they were ambiguous. However, we did not remove misspelled entities (typos) because they referred to specific (unambiguous) persons. Retaining them for the evaluations helped us test the effectiveness of our method. Because identifying different mentions of an entity correctly is difficult (Lee et al., 2013), we did not consider coreferences of a person name. We counted the frequency of each extracted person name and found that many of the names rarely occurred in the topic documents. The rank-frequency distribution of person names follows Zipf's law (Zipf, 1949). Low frequency names are usually persons that are irrelevant to the topic (e.g., journalists), so they were excluded from the evaluation. Thus, for the evaluation, we selected the first frequent person names whose accumulated frequencies reached  $\lambda$  percent of the total

<sup>2</sup> <http://nlp.stanford.edu/software/CRF-NER.shtml>



person name frequency count. In other words, the evaluated person names accounted for  $\lambda$  percent of the person names in the examined topic. In the following experiments, we assess the system performance under  $\lambda = 50\%$  and  $\lambda = 60\%$ . All the evaluated person names represent important topic persons.

We asked two experts to annotate the person stances and establish a reliable ground truth for the evaluations. Then, to evaluate the stance identification performance, we utilized the rand index (Rand, 1971), a conventional evaluation metric frequently used to compare clustering algorithms. More specifically, the rand index is based on person pairs. After a set of topic persons is partitioned into clusters, the rand index measures the percentage of clustering decisions that are correct (e.g., placing a person pair with the same stance in the same cluster). As large topics generally dominate small topics in micro-averaging (Manning & Schütze, 1999), we use macro-averaging to average the rand index scores of the evaluated topics. Paragraph tags are not provided in the evaluated topic documents, so a block is a topic document in our evaluation.

### 3.3.2 Effect of system components

In this section, we discuss the system parameter  $\gamma$ , and consider the effects of the weighted correlation coefficient, OBE, and the MaxMin initialization algorithm. Parameter  $\gamma$  is a similarity threshold that is used to eliminate dissimilar (off-topic) blocks. In the experiment, we set  $\gamma$  at 0.1 initially and increased the value in increments of 0.1 to 0.9. We did not consider  $\gamma = 0$  or 1 because the range of the cosine similarity is  $[0,1]$ . Thus, setting  $\gamma = 0$  would not remove any topic blocks; while setting  $\gamma = 1$  would eliminate all topic blocks so that the block set  $B$  would be empty and the stance identification process could not be implemented. To measure the true effect of  $\gamma$ , we excluded the influence of other system components. We did not utilize the MaxMin

initialization algorithm. In addition, we set the parameter  $\beta$  of the weighted correlation coefficient at 0.5; that is, we used the primitive correlation coefficient. For each setting of  $\gamma$ , we randomly initialized our method twenty times and averaged the results for comparison.

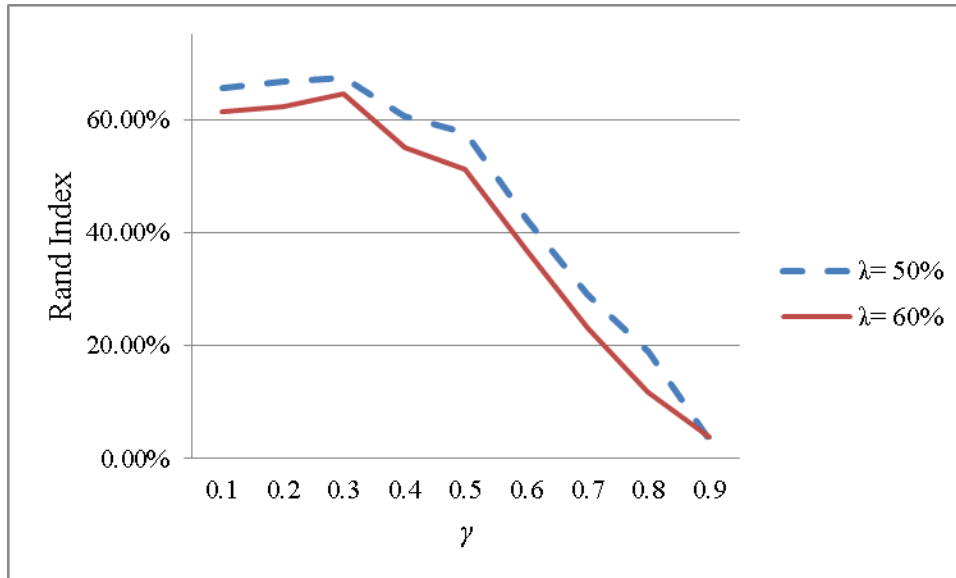
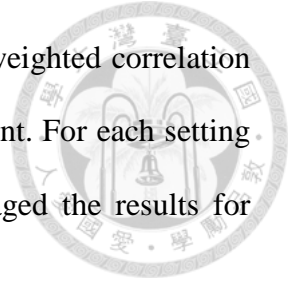


Figure 6. The rand index scores under different settings of  $\gamma$

Figure 6 shows the rand index scores under different settings of  $\gamma$ . The rand index decreases when  $\gamma$  is larger than 0.3. Recall that OBE removes blocks whose cosine similarity to the representative vector of a topic is lower than  $\gamma$ . As the topic representative vector is the average of the blocks  $\underline{B}$ , it covers the most frequent topic persons. Thus, a large  $\gamma$  setting excludes blocks that cover less frequent person names and thereby reduces the correlation between persons with the same stance. As a result, the stance identification performance deteriorates. For instance, in Topic  $A_{11}$ , Kevin Garnett, a franchise player with Celtics, and his teammates have an average correlation coefficient of 0.3773 under  $\gamma = 0.3$ . However, under  $\gamma = 0.4$ , the average correlation coefficient drops to 0.2887. This example demonstrates that OBE with a large  $\gamma$  setting

eliminates the blocks that cover the less frequent persons associated with Celtics and reduces the correlation between Kevin Garnett and his teammates. Because our method clusters topic persons in terms of the correlation coefficient, the lower correlation causes the method to cluster topic persons incorrectly. Overall, setting  $\gamma$  at 0.3 achieves the best stance identification performance. Therefore, we utilize the setting in subsequent evaluations.

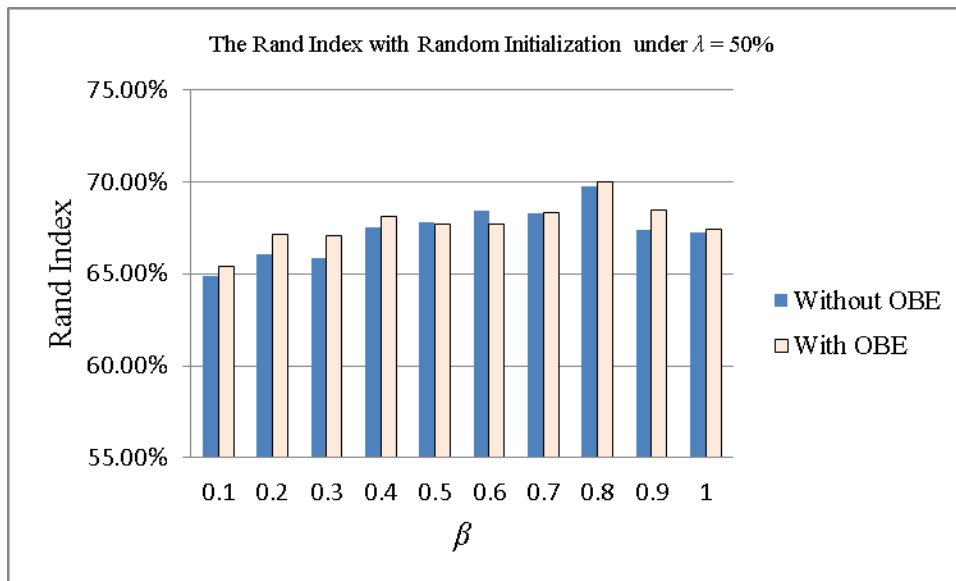


Figure 7. The rand index with random initialization under  $\lambda = 50\%$

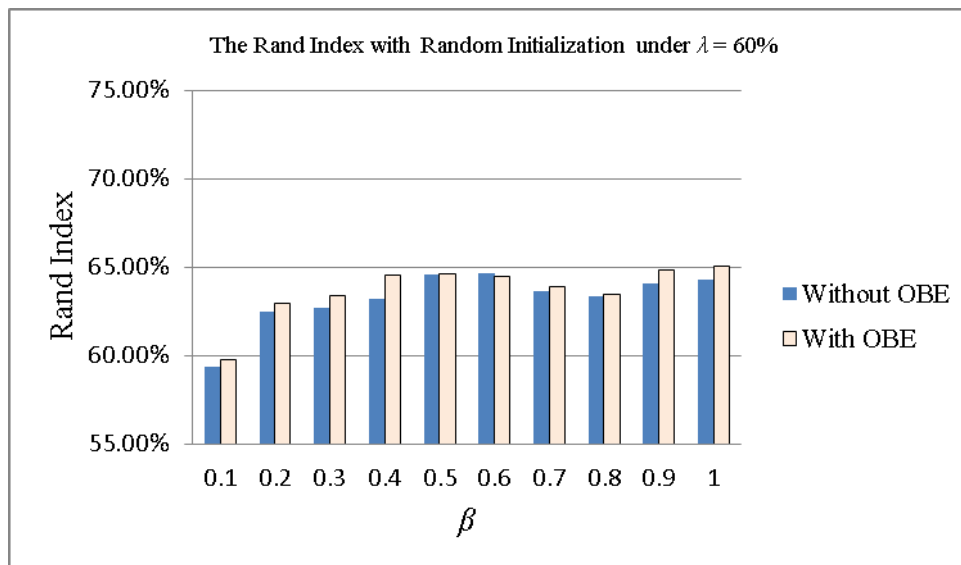
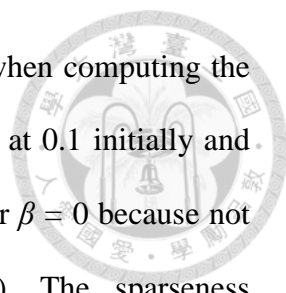


Figure 8. The rand index with random initialization under  $\lambda = 60\%$

The weighted correlation coefficient uses parameter  $\beta$  to adjust the weight of



non-co-occurring blocks and mitigate the text sparseness problem when computing the correlations between topic persons. To examine its effect, we set  $\beta$  at 0.1 initially and increased its value in increments of 0.1 up to 1. We did not consider  $\beta = 0$  because not all persons have co-occurring blocks (i.e.,  $co(i,k)$  is empty). The sparseness phenomenon leads to a zero denominator in Eq. (9) such that the weighted correlation coefficient is non-calculable. For each setting of  $\beta$ , we randomly initialize our method twenty times and average the stance identification results for comparison. We also compare the performance with and without OBE to examine the effect of off-topic blocks. When OBE is employed, parameter  $\gamma$  is set at 0.3 because of its superior performance in the previous experiment. Figures 7 and 8 show the performance under  $\lambda = 50\%$  and  $60\%$  respectively. The rand index scores under  $\lambda = 50\%$  are generally higher than those under  $\lambda = 60\%$ . This is because a large  $\lambda$  (i.e.,  $\lambda = 60\%$ ) includes infrequent topic persons, which make the stance identification task more difficult. As shown in the figures, using OBE generally improves the rand index scores. When collecting the experimental data, we found that some of the topic blocks were off-topic. As mentioned earlier, off-topic blocks would make uncorrelated persons, i.e., people with opposite viewpoints, positively correlated, and therefore have a negative impact on the stance identification performance. The blocks eliminated by OBE only account for 9.22% of the topic content, but their removal improves the correlation coefficient between topic persons. For example, in topic  $A_5$ , the original correlation coefficient between Pau Gasol and Rafer Alston, who play for Lakers and Magic respectively, is positive. However, after using OBE, they become negatively correlated with a coefficient of -0.0046. Since our method is based on the correlation coefficient, OBE improves the performance of stance identification. The improvement of OBE under  $\lambda = 60\%$  is slightly smaller than that under  $\lambda = 50\%$ . This is because  $\lambda = 60\%$  includes too many infrequent persons and

OBE only excludes a small portion of the topic content. As a result, OBE only corrects the correlations between some persons, so the performance improvement is smaller.

Next, we examine the effect of  $\beta$  and the weighted correlation coefficient. Figures 7 and 8 show that, generally, a small  $\beta$  setting yields an inferior performance. We find that many topic blocks, especially in the sports topics, are recaps of competing stances, which tend to mention persons of different stances together. As a small  $\beta$  value makes such co-occurring blocks important, the corresponding rand index score is lower. It is noteworthy that  $\beta = 1$  degrades the performance when  $\lambda = 50\%$ . Under this setting, the weighted correlation coefficient excludes all the co-occurring blocks and only uses the non-co-occurring blocks to determine the relationship between topic persons. However, the evaluated persons under  $\lambda = 50\%$  are so frequent that they appear in almost every topic block. Therefore, the weighted correlation coefficient is based on a few blocks that bias the relationship between topic persons; consequently, they have a negative impact on the stance identification performance. To summarize, by eliminating off-topic blocks, a large  $\beta$  setting usually yields a superior stance identification performance. The reason is that, when off-topic blocks are eliminated, the set of non-co-occurring blocks reveal either adverse relationships between persons or the absence of any relationships. Therefore, the stance identification performance improves as  $\beta$  increases. Moreover, this setting outperforms  $\beta = 0.5$  without OBE (i.e., the primitive EM approach). Hence, the proposed off-topic block elimination method and weighted correlation coefficient method reduce the text sparseness problem effectively.

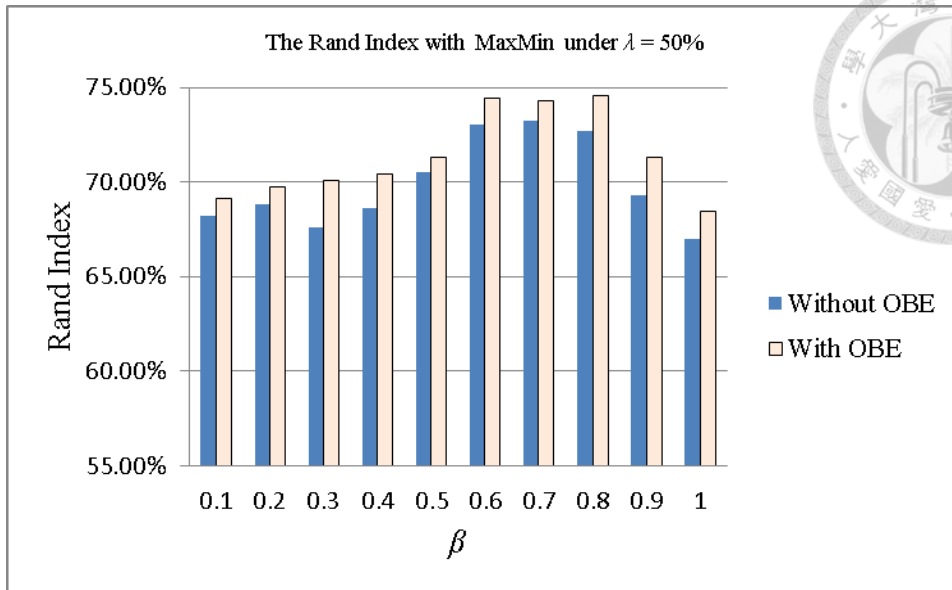


Figure 9. The rand index with MaxMin under  $\lambda = 50\%$

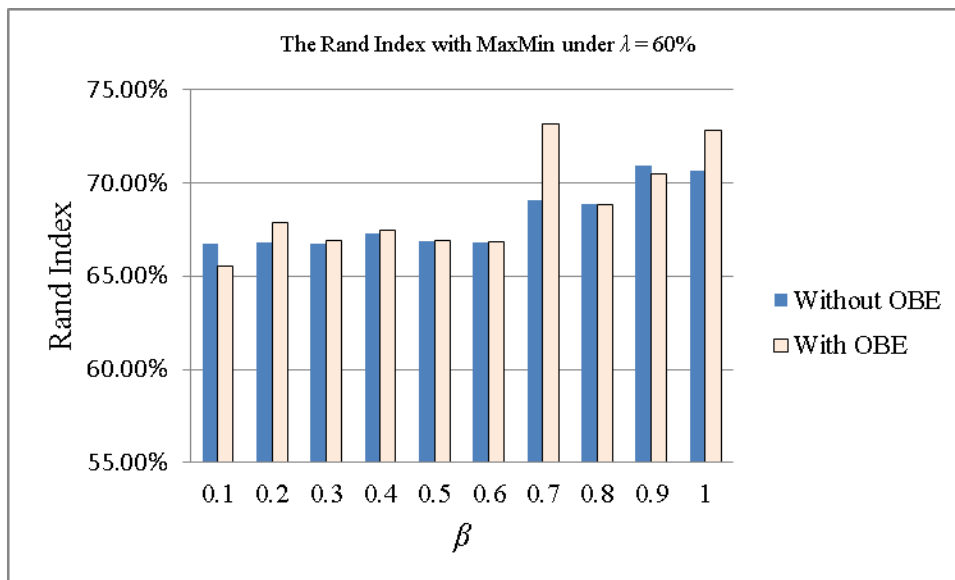
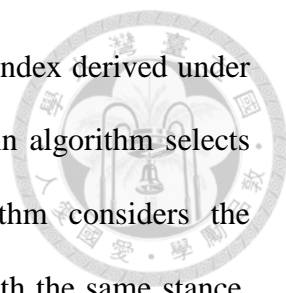


Figure 10. The rand index with MaxMin under  $\lambda = 60\%$

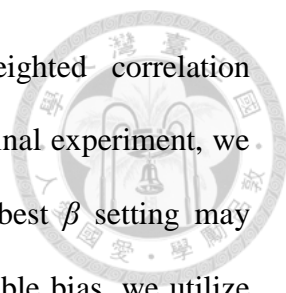
Finally, we consider the MaxMin initialization algorithm. Figures 9 and 10 show its effect under various parameter settings. Similar to the previous result, OBE improves the system performance and a large  $\beta$  increases the rand index score. Compared with the results in Figures 7 and 8, the rand index derived by using the MaxMin initialization algorithm is superior. The results demonstrate the importance of model initialization. As the performance of model-based EM methods is sensitive to model initialization



(Figueiredo & Jain, 2002; Pernkopf & Bouchaffra, 2005), the rand index derived under a random initialization strategy is inferior. By contrast, the MaxMin algorithm selects persons that are representative of stances. Because the algorithm considers the correlation between persons, it prevents the selection of persons with the same stance. Consequently, it enhances the representativeness of the initial stance model and thereby ensures a superior stance identification result.

### 3.3.3 Comparison with other clustering methods

Identifying the stances of topic persons is a special clustering problem that groups topic persons into stance-coherent clusters. Here, we compare our model-based EM method with the following well-known clustering methods; the K-means method (Manning & Schütze, 1999), the HAC method (Manning et al., 2008), the PLSI method (Hofmann, 1999), and the PCA-based method (Chen et al., 2010; 2012). Under K-means and HAC, we represent a topic person as a high-dimensional frequency vector (i.e.,  $p_i$ ) and use the cosine similarity to group similar persons into clusters. For HAC, we consider four well-known inter-cluster similarity strategies, namely, single-link, complete-link, average-link, and centroid-link strategies (Manning et al., 2008). For the PLSI method, a latent concept is represented by a variable  $z$  and the terms of a text corpus are clustered according to the probability  $P(z|w)$  (Hofmann, 1999). In our experiment,  $z$  is a stance and a term  $w$  is a person name. The PCA-based method also represents a topic person as a frequency vector. Because the method identifies topic persons' stances in terms of the sign of the entries in the principal eigenvector, it is only used to evaluate two-stance topics. In addition to the above methods, we compare a baseline method, which simply assumes that all topic persons have the same stance. The baseline comparison allows us to evaluate the efficiency of the clustering-based stance



identification methods. The proposed method utilizes the weighted correlation coefficient and OBE because of their superior performance. In the final experiment, we evaluate the effect of the  $\beta$  setting on all the topics. Using the best  $\beta$  setting may overestimate our system's performance. Therefore, to avoid a possible bias, we utilize the leave-one-out validation approach (Manning et al., 2008) to evaluate our method over multiple runs. In each run, a topic is selected for testing, and the remaining topics are used to derive the value of  $\beta$ . Then, the results of the evaluations of all the topics are averaged for comparison. As the clustering performances of K-means, PLSI, and our model-based EM method depend on cluster initialization, we initialize the methods randomly twenty times and select the best, average, and worst results for comparison. We also evaluate the proposed MaxMin initialization algorithm. To ensure that the comparisons are fair, each compared method partitions the evaluated topic persons of Topics  $A_1 \sim A_8$  into two stances. For Topics  $A_9 \sim A_{16}$ , the topic persons are clustered into four stances.

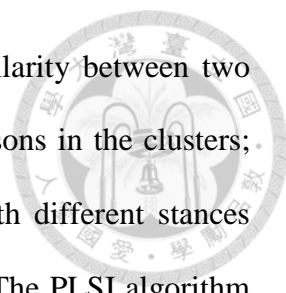
Table 2 shows the rand index scores of the compared methods. Because it is impossible to determine which initialization achieves the best performance, we compare the average performances of K-means, PLSI, and our method. As shown in the table, all the clustering-based methods outperform the baseline method. Intuitively, it is difficult to identify the person stances of Topics  $A_9 \sim A_{16}$  because they involve four stances and consider a lot of topic persons. For example, under  $\lambda = 60\%$ , there are 135 evaluated persons in Topics  $A_9 \sim A_{16}$ , but only 63 persons in the two-stance Topics  $A_1 \sim A_8$ . According to Zipf's Law, many person names rarely occur in the topic blocks. Their sparseness makes the stance identification of the four-stance topics more difficult. In our experiment, however, the rand index scores of the four-stance topics are higher than those of the two-stance topics because the number of two-stance persons is small.



Table 2. The stance identification results of the compared methods

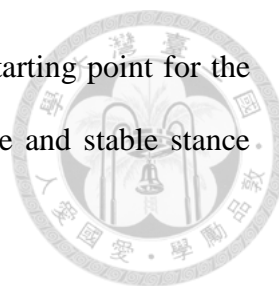
	Topics $A_1 \sim A_8$		Topics $A_9 \sim A_{16}$	
	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 50\%$	$\lambda = 60\%$
PCA-based method	61.91%	58.79%	n.a.	n.a.
Our method (MaxMin)	58.89%	67.39%	83.23%	73.12%
Our method (Random, best)	<b>77.19%</b>	<b>82.59%</b>	<b>89.72%</b>	<b>78.91%</b>
Our method (Random, avg.)	57.58%	59.96%	76.70%	69.48%
Our method (Random, worst)	46.67%	37.80%	63.29%	58.33%
K-means (best)	68.57%	81.10%	75.93%	73.13%
K-means (avg.)	52.75%	52.07%	65.03%	61.24%
K-means (worst)	44.29%	37.16%	54.89%	49.79%
PLSI (best)	74.99%	72.52%	80.48%	71.38%
PLSI (avg.)	53.92%	54.89%	69.36%	64.61%
PLSI (worst)	41.11%	42.06%	61.29%	57.33%
HAC (single-link)	47.94%	50.56%	70.13%	59.45%
HAC (complete-link)	48.73%	54.97%	73.45%	64.77%
HAC (average-link)	53.49%	54.11%	73.15%	65.86%
HAC (centroid-link)	48.73%	58.69%	70.50%	61.99%
Baseline	36.98%	32.23%	19.63%	22.97%

Therefore, an incorrectly identified person stance of the two-stance topics has a significant effect on the system performance such that the rand index score is low. Nevertheless, our method still outperforms the compared methods on the two-stance topics. As shown in the table, although the PCA-based method yields a superior stance identification performance, it cannot deal with the four-stance topics. The performance of the K-means method is inferior when popular persons are selected as the centroids of the initial clusters. A topic person is considered popular if his/her name appears in several topic blocks. The frequency vector of a popular person usually contains a lot of non-zero entries, which tend to produce a high cosine similarity score because the cosine similarity is the inner product of the normalized vectors. For instance, in Topic  $A_5$ , Kobe Bryant and Dwight Howard have a high similarity score because they are popular (franchise) players of Lakers and Magic respectively. Under K-means, selecting such a person as the centroid of the initial cluster would merge cosine-similar but stance-different persons, and therefore impact the stance identification performance. The inferior performance of the HAC single-link strategy also reflects the shortcomings



of the cosine similarity measure. The strategy determines the similarity between two clusters by examining the cosine similarity of the most similar persons in the clusters; hence, a high cosine similarity score between popular persons with different stances would result in the merging of groups that have opposing stances. The PLSI algorithm also groups popular but stance-different persons together because its objective function tends to compute a high  $P(z|w)$  for person names that co-occur frequently in topic blocks. By contrast, our method determines the relationships of persons in terms of the correlation coefficient, which shows how the occurrences of person names and stances vary jointly. Therefore, it can identify the relationships between popular persons correctly. For instance, the correlation coefficient between Kobe Bryant and Dwight Howard is -0.13, so our method achieves a better stance identification performance than the compared methods.

Finally, we assess the performance of the MaxMin initialization algorithm. As mentioned previously, EM methods are sensitive to model initialization, so an effective initialization algorithm is essential to ensure stable stance identification results. The MaxMin algorithm initializes our stance identification method by selecting representative persons of different stances. To prevent the selection of persons with the same stance, it considers the correlation coefficient between persons and selects those with low correlations. However, because of the text sparseness problem, the correlation coefficient is sometimes underestimated so that persons with the same stance are selected. For example, in the IMF topic, MaxMin selects Alain Juppe, Vladimir Putin, Angela Merkel, and Elena Salgado, but Alain Juppe and Angela Merkel have the same stance. As a result, the stance identification performance is inferior to the best result. Nevertheless, MaxMin produces comparable results and outperforms our average performance. Moreover, it outperforms the compared methods on difficult topics. The



results indicate that the MaxMin algorithm always selects a good starting point for the model search task, and that helps our EM method identify accurate and stable stance identification results.

### 3.3.4 Person stance identification examples

In this section, we consider two four-stance topics, namely, the 2008 NBA Conference Finals (Topic  $A_9$ ) and the 2009 NBA Conference Finals (Topic  $A_{10}$ ), to show that the proposed method can identify stance dynamics. Figure 11 shows the person stance identification results, and Tables 3 and 4 detail the expectation values of the topic persons. The first column in each table lists the evaluated persons of the topics and the remaining columns list the expectation  $E[h_{i,k}]$  generated by our method. A person belongs to the stance with the maximum expectation.

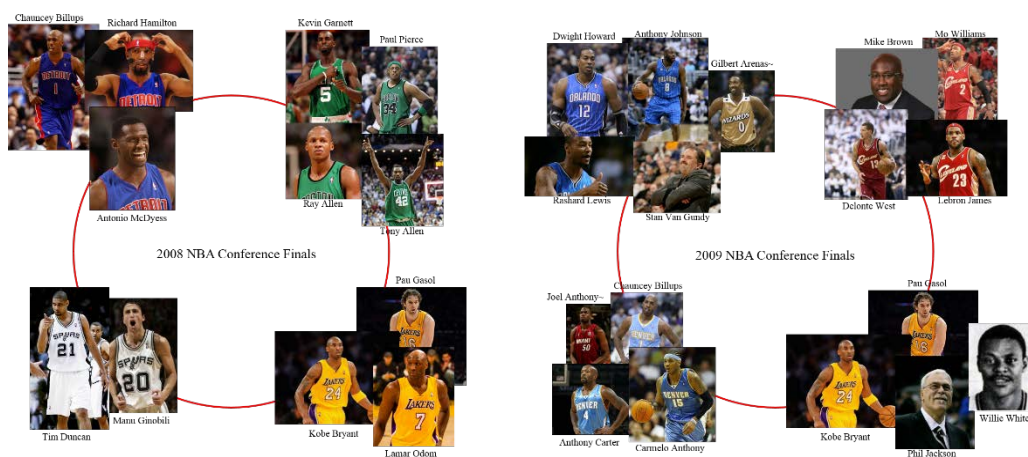


Figure 11. The person stance identification results of the 2008 and 2009 NBA Conference Finals.

In Figure 11, each group corresponds to a basketball team that competed in the NBA Conference Finals. The results show that the proposed method identifies the

stances of the important persons in the topics accurately. In fact, for Topic  $A_9$  (the left hand side of Figure 11), the important players in the conference finals are grouped perfectly. For Topic  $A_{10}$ , our method incorrectly identifies three neutral players (marked by ~) who were not involved in the 2009 NBA Conference Finals, but were mentioned frequently in comparison to other players. Even so, if we ignore the neutral entities, which are always wrong, irrespective of the stance identification method employed, our method identifies important players perfectly. It is noteworthy that, in Topic  $A_9$ , our method correctly identifies Chauncey Billups as a member of the Detroit Pistons, as shown in Figure 11. However, after the 2008 NBA season, Billups was traded to the Denver Nuggets. As our method identifies person stances in terms of word usage patterns in topic documents, it captured the stance dynamics and identified Billups correctly as a member of the Denver Nuggets. The examples demonstrate that our unsupervised method is context-oriented and can identify stance dynamics without using any external knowledge source.

Table 3. The stance identification results for Topic  $A_9$  ( $\lambda = 60\%$ ,  $\beta = 0.4$  with OBE)

	$E[h_{i,Pistons}]$	$E[h_{i,Celtics}]$	$E[h_{i,Spurs}]$	$E[h_{i,Lakers}]$
Chauncey Billups	<b>0.41</b>	0.28	0.14	0.17
Richard Hamilton	<b>0.36</b>	0.29	0.16	0.19
Antonio McDyess	<b>0.47</b>	0.14	0.15	0.24
Kevin Garnett	0.30	<b>0.36</b>	0.16	0.18
Paul Pierce	0.32	<b>0.33</b>	0.16	0.19
Ray Allen	0.30	<b>0.39</b>	0.14	0.17
Tony Allen	0.25	<b>0.37</b>	0.18	0.20
Manu Ginobili	0.19	0.17	<b>0.37</b>	0.27
Tim Duncan	0.17	0.16	<b>0.38</b>	0.29
Kobe Bryant	0.16	0.15	0.33	<b>0.36</b>
Lamar Odom	0.17	0.16	0.23	<b>0.44</b>
Pau Gasol	0.19	0.17	0.31	<b>0.33</b>

Table 4. The stance identification results for Topic  $A_{10}$  ( $\lambda = 60\%$ ,  $\beta = 0.4$  with OBE)

	$E[h_{i, Magic}]$	$E[h_{i, Lakers}]$	$E[h_{i, Cav}]$	$E[h_{i, Nugget}]$
Anthony Johnson	<b>0.36</b>	0.24	0.20	0.20
Dwight Howard	<b>0.31</b>	0.19	0.30	0.20
Rashard Lewis	<b>0.42</b>	0.17	0.24	0.17
Stan Van Gundy	<b>0.30</b>	0.20	0.29	0.21
Gilbert Arenas~	<b>0.30</b>	0.20	0.28	0.22
Kobe Bryant	0.18	<b>0.42</b>	0.17	0.23
Pau Gasol	0.21	<b>0.37</b>	0.20	0.22
Phil Jackson	0.22	<b>0.35</b>	0.21	0.22
Willie White~	0.25	<b>0.27</b>	0.25	0.23
Delonte West	0.25	0.17	<b>0.41</b>	0.17
Lebron James	0.31	0.17	<b>0.34</b>	0.18
Mike Brown	0.25	0.22	<b>0.32</b>	0.21
Mo Williams	0.31	0.18	<b>0.32</b>	0.19
Anthony Carter	0.19	0.26	0.19	<b>0.36</b>
Carmelo Anthony	0.19	0.27	0.18	<b>0.36</b>
Chauncey Billups	0.21	0.27	0.22	<b>0.30</b>
Joel Anthony~	0.19	0.25	0.19	<b>0.37</b>

### 3.4 Conclusions of the EM method

We proposed an effective EM method for identifying person stances in topics without using external knowledge sources. To solve the off-topic block and text sparseness problems, we incorporate two techniques into our EM method. The experiment results demonstrate that the techniques can solve the problems effectively. As the EM method is sensitive to model initialization, we propose the MaxMin initialization algorithm which yields stable and accurate stance identification results. The proposed stance identification method is unsupervised, so it can be applied to different domains and can capture the stance dynamics without using any external knowledge source.

#### 4. A topic person stance identification method based on friendship network analysis

In our first approach, we didn't consider the competing semantics of documents and didn't employ the features of social network. We also observed that few of topic persons are stance-irrelevant, and that affected the performance of topic person stance identification. Hence, we proposed a stance identification method, SCIFNET, which groups the persons mentioned in topic documents into stance-coherent clusters, to cope with the problems. Figure 12 shows SCIFNET's system architecture, which is comprised of three components: *friendship network construction*, *stance community expansion*, and *stance community refinement*. Specifically, given a set of documents reporting a topic with  $K$  stances, SCIFNET first extracts the topic persons mentioned in the documents. Then, it constructs a friendship network of the topic persons based on the co-occurrence of the persons in the documents and the stance orientation of the documents. Next, the stance community expansion process considers the stance identification of topic persons as a community detection task and iteratively expands the  $K$  stances (i.e., communities) in the friendship network. In the last phase, the stance community refinement algorithm improves the stance identification result in accordance with an objective function, which measures the stance coherence of the detected communities. Note that an issue in community detection is to determine the number of communities in a network. Like many community detection methods, e.g., (Ding et al., 2001), (Yang et al., 2009), and (Gao et al., 2010), we assume that the number of communities (i.e.,  $K$ ) is known in advance. In the following subsections, we describe each system component in detail. We also show that using the components increases the value of the objective function such that the stance identification result converges to a local optimum.

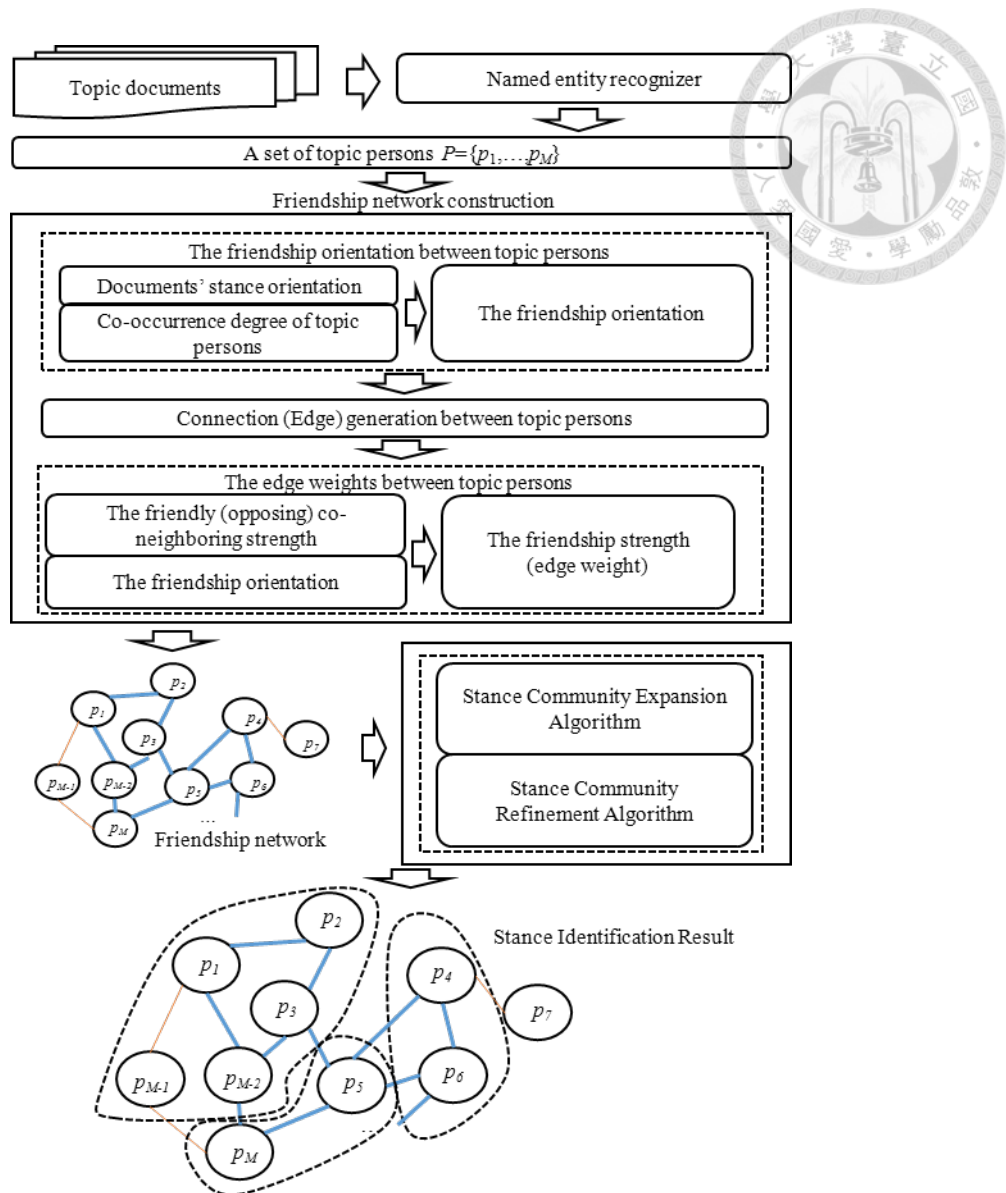


Figure 12. The system architecture

#### 4.1 Friendship network construction

Let  $D = \{d_1, d_2, \dots, d_N\}$  be a set of topic documents, and let  $P = \{p_1, p_2, \dots, p_M\}$  be a set of topic persons mentioned in  $D$ . The friendship network construction generates a friendship network  $G = \{P, E\}$ , where the topic persons in  $P$  form the network's nodes; and  $E = \{(p_i, p_j)\}$  is a set of edges that indicate the friendship orientation of the topic persons (i.e., whether the association between the persons is friendly or opposing). Generally, it is difficult to discover friendship orientations from text. However, Harris (1954) observed that text units with opposing meanings seldom co-occur in the same

context. In addition, Kanayama and Nasukawa (2006) showed that text units with the same sentiment tend to occur (not occur) jointly to make the contexts coherent. Hence, the correlation coefficient (Keller, 2008), which measures the co-occurrence degree of topic persons in  $D$ , is probably a good measure for discovering the friendship orientation between topic persons. Nevertheless, we found that topic documents sometimes cover controversial issues. In the documents, people with different stances strongly criticize each other. Thus, only considering the co-occurrence degree of topic persons in  $D$  may overestimate the friendship of rivals and degrade the performance of topic person stance identification. Intuitively, topic persons who frequently co-occur in stance-friendly (stance-opposing) documents may have a friendly (opposing) association. To quantitate the stance orientation of a topic document, we adopt Turney and Littman (2003)'s method and compute the stance weight of a document as follows:

$$sw_d = \sum_{word_i \in d} \log \left( \frac{\prod_{word_j \in Fwords} count(word_i, word_j) \cdot \prod_{word_k \in Owords} count(word_k)}{\prod_{word_j \in Fwords} count(word_j) \cdot \prod_{word_k \in Owords} count(word_i, word_k)} \right), \quad (12)$$

where  $sw_d$  represents the stance weight of document  $d$ ; and  $Fwords$  and  $Owords$  are, respectively, sets of words with stance-friendly and stance-opposing semantics compiled by linguistic experts. The function  $count(word_i, word_j)$  returns the number of documents in which  $word_i$  and  $word_j$  co-occur in our topic corpus. Basically, the equation utilizes pointwise mutual information (PMI) to compute the stance weight of a document. The stance weight  $sw_d$  is positive if  $d$ 's content is strongly associated with  $Fwords$ , and negative if the content is strongly associated with  $Owords$ . We define the following stance-oriented correlation coefficient (SOCOR), which incorporates the stance weight into the correlation coefficient:





$$\begin{aligned}
socor(p_i, p_j) = & \frac{\sum_{d \in D_{friendly}} sw_d * (p_{i,d} - \bar{p}_{i,friendly}) * (p_{j,d} - \bar{p}_{j,friendly}) + \sum_{d \in D_{opposing}} sw_d * (p_{i,d} - \bar{p}_{i,opposing}) * (p_{j,d} - \bar{p}_{j,opposing})}{\sqrt{\sum_{d \in D_{friendly}} [\sqrt{sw_d} * (p_{i,d} - \bar{p}_{i,friendly})]^2 + \sum_{d \in D_{opposing}} [\sqrt{|sw_d|} * (p_{i,d} - \bar{p}_{i,opposing})]^2} * \sqrt{\sum_{d \in D_{friendly}} [\sqrt{sw_d} * (p_{j,d} - \bar{p}_{j,friendly})]^2 + \sum_{d \in D_{opposing}} [\sqrt{|sw_d|} * (p_{j,d} - \bar{p}_{j,opposing})]^2}}, \tag{13}
\end{aligned}$$

where  $D_{friendly} \subseteq D$  is a set of topic documents whose stance weight is positive;  $D_{opposing} \subseteq D$  is a set of topic documents whose stance weight is negative; and  $\bar{p}_{i,friendly}$  and  $\bar{p}_{i,opposing}$  are the average frequencies of  $p_i$  occurring in  $D_{friendly}$  and  $D_{opposing}$  respectively. Like the correlation coefficient, the range of  $socor(p_i, p_j)$  is within  $[-1, 1]$ . It is zero if the occurrences of  $p_i$  and  $p_j$  in  $D$  are independent of each other. However, if  $p_i$  and  $p_j$  tend to co-occur in stance-friendly (stance-opposing) documents, the  $socor(p_i, p_j)$  is positive (resp. negative). Next, we define the friendship orientation in terms of the stance-oriented correlation coefficient.

*Definition 1 - Friendship Orientation:*

*The friendship orientation between  $p_i$  and  $p_j$  is denoted by  $socor(p_i, p_j)$  and  $-1 \leq socor(p_i, p_j) \leq 1$ .*

We utilize SOCOR to construct the edge set  $E$ . In addition, to consolidate relationships between topic persons, we define a friendship orientation threshold  $\theta$ . An edge  $(p_i, p_j)$  is established if  $socor(p_i, p_j) > \theta$  or  $socor(p_i, p_j) < -\theta$ .

Jeh and Wisdom (2002) and Antonellis et al. (2008) demonstrated that the association between nodes in a network is proportional to their co-neighboring level. In other words, the greater the overlap between the neighbors of two nodes, the higher will be the likelihood that the nodes are associated with each other. In our research, however,



edges indicate either friendly orientations or opposing orientations. To measure the co-neighboring strength, we define two types of neighbors, namely, friendly neighbors and opposing neighbors.

*Definition 2 - Friendly Neighbors:*

Let  $p_i \in P$ . The friendly neighbors of  $p_i$ , denoted by  $\Gamma^+(p_i)$ , form a set of nodes whose friendship orientations to  $p_i$  are larger than  $\theta$ . Formally,

$$\Gamma^+(p_i) = \{p_j \in P \mid \text{socor}(p_i, p_j) > \theta\}.$$

*Definition 3 - Opposing Neighbors:*

Let  $p_i \in P$ . The opposing neighbors of  $p_i$ , denoted by  $\Gamma^-(p_i)$ , form a set of nodes whose friendship orientations to  $p_i$  are smaller than  $-\theta$ . Formally,

$$\Gamma^-(p_i) = \{p_j \in P \mid \text{socor}(p_i, p_j) < -\theta\}.$$

In Definitions 4 and 5, we employ the Jaccard coefficient to measure the friendly co-neighboring strength and the opposing co-neighboring strength respectively.

*Definition 4 - Friendly Co-neighboring Strength:*

The friendly co-neighboring strength between  $p_i$  and  $p_j$  is denoted by  $\gamma(p_i, p_j)$ :

$$\gamma(p_i, p_j) = \frac{|\Gamma^+(p_i) \cap \Gamma^+(p_j)|}{|\Gamma^+(p_i) \cup \Gamma^+(p_j)|}.$$

*Definition 5 - Opposing Co-neighboring Strength:*

The opposing co-neighboring strength between  $p_i$  and  $p_j$  is denoted by  $\omega(p_i, p_j)$ :



$$\omega(p_i, p_j) = \frac{|\Gamma^-(p_i) \cap \Gamma^-(p_j)|}{|\Gamma^-(p_i) \cup \Gamma^-(p_j)|}$$

Clearly, if two nodes share several friendly (opposing) neighbors, their friendly (opposing) co-neighboring strength is strong. Finally, we combine the friendship orientation with the co-neighboring strengths, and define the friendship strength, i.e., the edge weight, as follows.

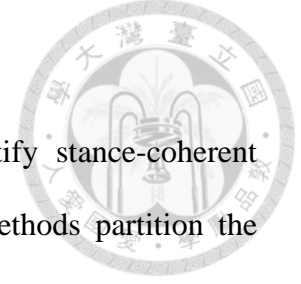
*Definition 6 - Friendship Strength:*

*The friendship strength, denoted by  $\delta(p_i, p_j)$ , represents the weight of edge  $(p_i, p_j)$ .*

$$\delta(p_i, p_j) = (socor(p_i, p_j) + 1)^{((\gamma(p_i, p_j) + \omega(p_i, p_j)) / 2 + \beta^\wedge)}, \text{ if } socor(p_i, p_j) > \theta.$$

$$\delta(p_i, p_j) = -(|socor(p_i, p_j)| + 1)^{(1 - ((\gamma(p_i, p_j) + \omega(p_i, p_j)) / 2 + \beta^\wedge))}, \text{ if } socor(p_i, p_j) < -\theta.$$

For friendly orientations (i.e.,  $socor(p_i, p_j) > \theta$ ), the friendly and opposing co-neighboring strengths function as an exponent to amplify the friendly relationships between nodes. We utilize a parameter  $\beta^\wedge \geq 1$  to ensure that the exponent is not less than 1; and we add 1 to a friendly orientation so that the base is greater than 1. As the enemies of foes may be friends, the friendship strength of  $p_i$  and  $p_j$  is strong and positive if they have a friendly orientation and share a lot of friendly and opposing neighbors. If  $p_i$  and  $p_j$  have an opposing orientation (i.e.,  $socor(p_i, p_j) < -\theta$ ), their friendship strength is negative. However,  $p_i$  and  $p_j$  may not fight against each other if they have many friends and adversaries in common. The negative friendship strength is thus diminished if the friendly and opposing neighbors of  $p_i$  and  $p_j$  overlap a great deal.



## 4.2 The objective function of SCIFNET

After constructing the friendship network of a topic, we identify stance-coherent communities in the network. In general, community detection methods partition the nodes of a network into clusters (i.e., communities) in accordance with the principle that maximizes the association between the nodes in each cluster, while minimizing the association between the clusters (Shi & Malik, 2000). We define the following objective function to identify a coherent stance identification result.

$$C = \arg \max_{\langle c_1, c_2, \dots, c_K \rangle} \sum_{cluster:m}^K \left[ \sum_{p_i, p_j \in c_m, i < j, (p_i, p_j) \in E} \delta(p_i, p_j) \right] - \sum_{cluster:m, n, m < n}^K \left[ \sum_{p_i \in c_m, p_j \in c_n, (p_i, p_j) \in E} \delta(p_i, p_j) \right], \quad (14)$$

where  $K$  is the number of clusters.  $\langle c_1, c_2, \dots, c_K \rangle$  is a set of stance clusters in which  $c_m \subseteq P$  and  $c_m \cap c_n = \text{null}$  for  $m \neq n$ . They provide a stance identification result. To maximize the objective function, a stance identification result needs to maximize the first term of Eq. (14) and minimize the second term simultaneously. In other words, the topic person stance identification method seeks a set of stance clusters that maximize the friendship strength within clusters (the first term of the objective function) and minimize the friendship strength between clusters (the objective function's second term).

## 4.3 Stance community expansion

Figure 13 shows the proposed stance community expansion algorithm, and Figure 14 provides an example of stance community expansion. In the algorithm, the symbol  $P_{unlabeled}$  represents a set of unlabeled nodes (i.e., topic persons). Initially,  $P_{unlabeled} = P$ ; that is, all nodes are unlabeled. The algorithm randomly selects  $K$  nodes as the seeds of stance clusters and expands the clusters iteratively by merging unlabeled nodes. In each

iteration, a set of unlabeled nodes  $U$  that connect directly to a stance cluster are identified (i.e.,  $U = \{p_i \in P_{unlabeled} \mid (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq K\}$ ). Each node  $p_i$  in  $U$  is then examined to determine an appropriate cluster label for it. Let  $Z_i$  denote the set of stance clusters that the unlabeled node  $p_i$  is connected to directly; that is,  $Z_i = \{c_k \mid (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq K\}$ . For instance,  $Z_4$  shown in Figure 14 comprises clusters  $c_1$  and  $c_2$ . We compute the merging score for each of the stance clusters  $c_k$  in  $Z_i$  as follows:

$$ms_{i,k} = \sum_{p_j \in c_k, (p_i, p_j) \in E} \delta(p_i, p_j), \quad (15)$$

*The Stance Community Expansion Algorithm:*

```

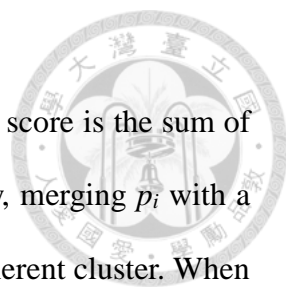
P_unlabeled = P
randomly select K nodes from P_unlabeled to form the seeds of {c_1, c_2, ..., c_K}
havePositiveMergingScore = true

while ( P_unlabeled ≠ ∅ & havePositiveMergingScore) do
    havePositiveMergingScore = false
    U = { p_i ∈ P_unlabeled | (p_i, p_j) ∈ E, p_j ∈ c_k, 1 ≤ k ≤ K }
    for each p_i in U do
        Z_i = { c_k | (p_i, p_j) ∈ E, p_j ∈ c_k, 1 ≤ k ≤ K }
        score_max = max_{c_k ∈ Z_i} ms_{i,k}
        cluster_max = arg max_{c_k ∈ Z_i} ms_{i,k}

        if score_max > 0 then
            c_{cluster_max} = c_{cluster_max} ∪ {p_i}
            P_unlabeled = P_unlabeled \ {p_i}
            havePositiveMergingScore = true
        end if
    end for
end while
return C = {c_1, c_2, ..., c_K}

```

Figure 13. The stance community expansion algorithm



where  $ms_{i,k}$  is the score of merging  $p_i$  with  $c_k$ . Basically, the merging score is the sum of the edge weights associated with  $p_i$  and stance cluster  $c_k$ . Intuitively, merging  $p_i$  with a cluster that has a positive merging score should produce a stance-coherent cluster. When more than one cluster has a positive merging score, the algorithm merges  $p_i$  with the stance cluster that has the maximum merging score. Below, we show that the step provides the most benefit for the objective function. Note that the merging score is negative if most of the nodes in  $c_k$  have an opposing friendship to  $p_i$ . Because merging  $p_i$  with a stance-opposing cluster is inappropriate, the algorithm revokes the merge operation if the maximum merging score is negative. The algorithm iteratively expands stance clusters until all the unlabeled nodes in the friendship network are merged or no unlabeled node has a positive merging score with any stance cluster. Then, it returns a stance identification result which will be polished by the stance community refinement algorithm.

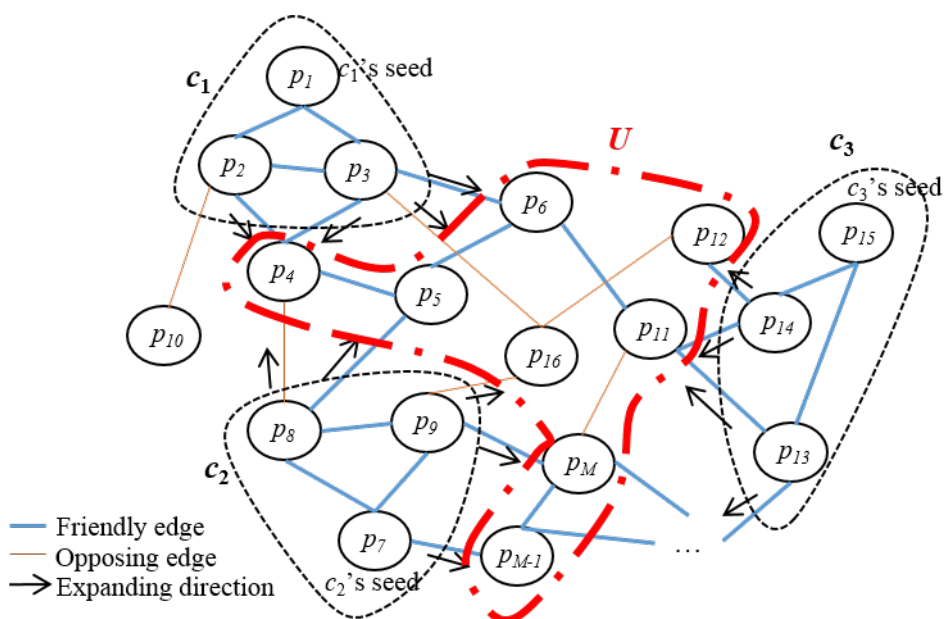
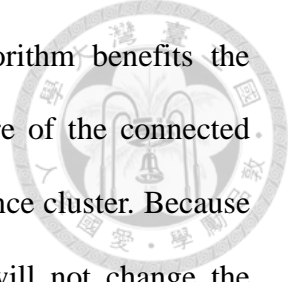


Figure 14. An example of stance community expansion



The following cases show how the merge step of the algorithm benefits the objective function. In the first case,  $|Z_i| = 1$  and the merging score of the connected stance cluster is positive<sup>3</sup>. Here,  $p_i$  is merged with the connected stance cluster. Because there is no other connected stance cluster, the merge operation will not change the second term of the objective function. Moreover, the operation increases the first term of the objective function by the positive merging score, so it benefits the objective function. In the second case,  $|Z_i| > 1$  and the maximum merging score is positive<sup>4</sup>. Next, we show that merging  $p_i$  with the stance cluster that has the maximum merging score provides the most benefit for the objective function.

*Proof:*

Let  $|Z_i| = k$ , and let  $k > 1$ . We have a sequence of merging scores  $\langle ms_{i,1}, ms_{i,2}, \dots, ms_{i,k} \rangle$  for the stance clusters in  $Z_i$ . Let  $ms_{i,1} \geq ms_{i,2} \geq \dots \geq ms_{i,k}$  and let  $ms_{i,1} > 0$ . The stance community expansion algorithm merges  $p_i$  with  $c_1$ . The inequality  $ms_{i,1} \geq ms_{i,n}$  holds for any stance cluster  $c_n$  in  $Z_i$  if  $n \neq 1$ . In other words,

$$\sum_{p_j \in c_1} \delta(p_i, p_j) \geq \sum_{p_j \in c_n} \delta(p_i, p_j). \quad (16)$$

Because  $Z_i$  has been determined, the summation of  $\langle ms_{i,1}, ms_{i,2}, \dots, ms_{i,k} \rangle$  (i.e.,  $\sum_{l=1 \text{ to } k} ms_{i,l}$ ) is a fixed value. The inequality  $ms_{i,1} \geq ms_{i,n}$  also implies that

$$\sum_{l \neq 1} ms_{i,l} \leq \sum_{l \neq n} ms_{i,l} \quad (17)$$

<sup>3</sup> We exclude the case where  $|Z_i| = 1$  and the merging score is negative. This is because the algorithm will not merge  $p_i$  with any stance cluster.

<sup>4</sup> We exclude the case where the maximum merging score is negative because the algorithm will not merge  $p_i$  with any stance cluster.



That is,

$$\sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) \leq \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j) \quad (18)$$

or

$$- \sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) \geq - \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j). \quad (19)$$

By combining Equations (16) and (19), we have

$$\begin{aligned} \sum_{p_j \in c_1} \delta(p_i, p_j) - \sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) &\geq \\ \sum_{p_j \in c_n} \delta(p_i, p_j) - \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j). & \end{aligned} \quad (20)$$

The above inequality indicates that if the unlabeled node  $p_i$  is associated with more than one stance cluster, the stance community expansion algorithm will merge  $p_i$  with the cluster that benefits the objective function the most.

□



#### 4.4 Stance community refinement

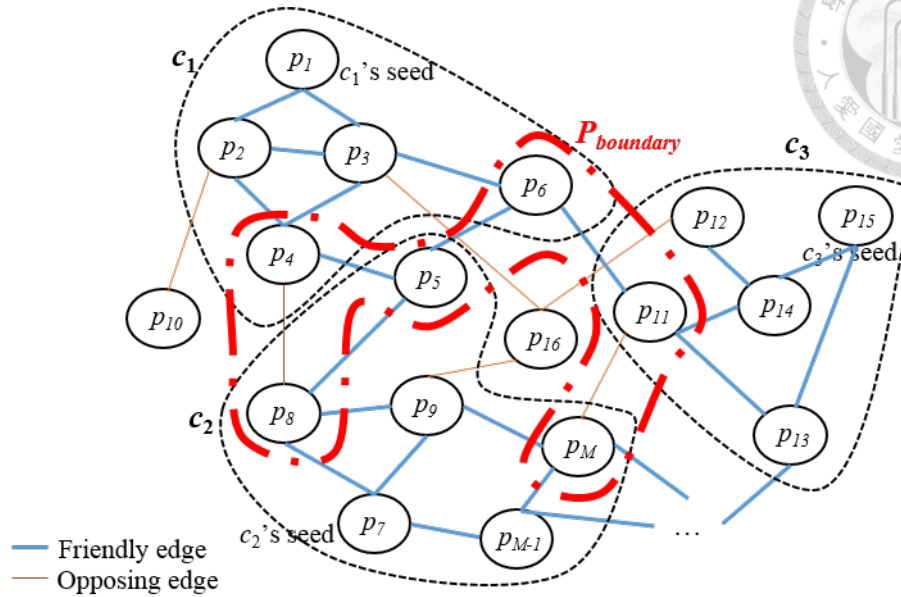


Figure 15. An example of stance community refinement

The stance community expansion algorithm iteratively expands stance clusters from the seed nodes. In some cases, a node is merged with a stance cluster simply because it is close to the cluster's seed. However, it may be better to merge the node with some other cluster. For instance, node  $p_5$  in Figure 15 is merged with cluster  $c_2$  even though it is strongly associated with cluster  $c_1$ . To minimize the effect of this “early merging” problem, we developed the following stance community refinement algorithm. The algorithm refines the clusters iteratively. In each iteration, it identifies a set of boundary nodes  $P_{boundary} \subseteq P$ . Each node in  $P_{boundary}$  belongs to a stance cluster and also connects to some other stance clusters. In other words,  $P_{boundary} = \{p_i | (p_i, p_j) \in E, p_i \in c_m, p_j \in c_n, m \neq n\}$ . If there is no boundary node, the stance community refinement stops; otherwise, the algorithm re-clusters each boundary node to the stance cluster that produces the maximum merging score. The algorithm continues to identify and cluster boundary nodes until the clustering result is stable; that is, the value of the objective function converges to a local optimum.



```

input:  $C = \{c_1, c_2, \dots, c_K\}$ 
 $C_{old} = \{\}$ 
while ( $C \neq C_{old}$ ) do
     $C_{old} = C$ 
     $P_{boundary} = \{p_i | (p_i, p_j) \in E, p_i \in c_m, p_j \in c_n, m \neq n\}$ 
    if  $P_{boundary} = \phi$  then
        break
    end if
    for each  $p_i$  in  $P_{boundary}$  do
         $c_{original} =$  the cluster that  $p_i$  belongs to
         $Z_i = \{c_k | (p_i, p_j) \in E, p_j \in c_k, 1 \leq k \leq K\}$ 
         $score_{max} = \max_{c_k \in Z_i} ms_{i,k}$ 
         $cluster_{max} = \arg \max_{c_k \in Z_i} ms_{i,k}$ 
        if  $c_{cluster_{max}} \neq c_{original}$  then
             $c_{cluster_{max}} = c_{cluster_{max}} \cup \{p_i\}$ 
             $c_{original} = c_{original} \setminus \{p_i\}$ 
             $P_{boundary} = P_{boundary} \setminus \{p_i\}$ 
        end if
    end for
end while
return  $C$ 

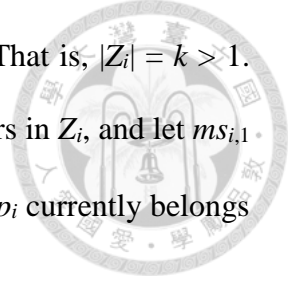
```

Figure 16. The stance community refinement algorithm

The core task of the stance community refinement algorithm is boundary node re-clustering. To demonstrate the convergence of the algorithm, we prove that the value of the objective function increases monotonically in each boundary node re-clustering operation.

*Proof:*

Let  $p_i$  be a boundary node. As a boundary node belongs to a stance cluster and also



connects to some other stance clusters,  $|Z_i|$  must be greater than 1. That is,  $|Z_i| = k > 1$ . Let  $\langle ms_{i,1}, ms_{i,2}, \dots, ms_{i,k} \rangle$  be the merging scores of the stance clusters in  $Z_i$ , and let  $ms_{i,1} \geq ms_{i,2} \geq \dots \geq ms_{i,k}$ . In addition, let  $c_n \in Z_i$  be the stance cluster that  $p_i$  currently belongs to. The inequality  $ms_{i,1} \geq ms_{i,n}$  holds. In other words,

$$\sum_{p_j \in c_1} \delta(p_i, p_j) \geq \sum_{p_j \in c_n} \delta(p_i, p_j). \quad (21)$$

Because  $Z_i$  has been determined, the summation of  $\langle ms_{i,1}, ms_{i,2}, \dots, ms_{i,k} \rangle$  (i.e.,  $\sum_{l=1 \text{ to } k} ms_{i,l}$ ) is a fixed value. The inequality  $ms_{i,1} \geq ms_{i,n}$  also implies that

$$\sum_{l \neq 1} ms_{i,l} \leq \sum_{l \neq n} ms_{i,l} \quad (22)$$

That is,

$$\sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) \leq \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j) \quad (23)$$

or

$$- \sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) \geq - \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j). \quad (24)$$

By combining Equations. (21) and (24), we have



$$\sum_{p_j \in c_1} \delta(p_i, p_j) - \sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) \geq \sum_{p_j \in c_n} \delta(p_i, p_j) - \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j) \quad (25)$$

Similar to the proof of stance community expansion, the above inequality indicates that the stance community refinement always re-clusters  $p_i$  into the cluster that benefits the objective function the most. The inequality also implies that

$$\sum_{p_j \in c_1} \delta(p_i, p_j) - \sum_{l=2 \text{ to } k} \sum_{p_j \in c_l} \delta(p_i, p_j) - \left[ \sum_{p_j \in c_n} \delta(p_i, p_j) - \sum_{l=1 \text{ to } k, l \neq n} \sum_{p_j \in c_l} \delta(p_i, p_j) \right] \geq 0 \quad (26)$$

The left-hand side of the inequality is equivalent to the variation in the objective function when  $p_i$  is re-clustered. Note that the variation is always non-negative. In other words, re-clustering the boundary nodes in  $P_{boundary}$  increases the value of the objective function monotonically. Because the set of possible stance identification results is finite, the stance community refinement algorithm will eventually find a local optimum.

□

#### 4.5 Stance-irrelevant topic person detection

A person mentioned in topic documents may be irrelevant to the topic stances. For instance, in the topic about the 2012 French Presidential Election, U.S. President Barack Obama, one of the most influential people in the world, was frequently mentioned in the topic documents because journalists liked to analyze his attitude toward the candidates. However, President Obama wasn't involved with the campaign and showed no

preference to any camp. SCIFNET can detect stance-irrelevant topic persons, which are defined as follows.

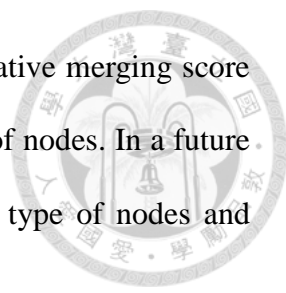


*Definition 7 - Stance-irrelevant Topic Persons:*

*Stance-irrelevant topic persons form a set  $P_{irrelevant} = \{p_i \in P \mid p_i \notin c_k, 1 \leq k \leq K\}$ .*

In other words, a topic person is stance-irrelevant if he/she does not belong to any stance cluster. SCIFNET classifies two types of nodes as stance-irrelevant because they cannot be merged with a stance cluster. The first is the set of outliers which have no connections to other nodes in a network (Xu et al., 2007). The nodes are stance-irrelevant because they do not show connections with any stance cluster. The second type comprises nodes that have connections with stance clusters; however, most of the connections are with clusters that have opposing associations with the nodes. Because the merging scores of the connected clusters are negative, the nodes cannot merge with any stance cluster.

Technically, we can increase the value of the objective function by merging a node that belongs to the second type with a cluster that does not have any connections with the node. For instance, merging node  $p_{10}$  in Figure 17 with  $c_2$  increases the value of objective function by 1.5. Even if the node connects to every stance cluster, the value of the objective function can still be increased by merging the node with the cluster that has the minimum negative merging score. For example, merging node  $p_{16}$  in Figure 17 with  $c_1$  increases the objective function value by 2.1. The above strategies increase the value of the objective function because they reduce the friendship strength between stance clusters, i.e., the second term of the objective function. However, although the two strategies are mathematically correct, merging a node with a cluster that does not



have any connections or with the cluster that has the minimum negative merging score is irrational. Hence, in this study, we do not merge the second type of nodes. In a future work, we will incorporate other information to handle the second type of nodes and refine the detection of stance-irrelevant topic persons.

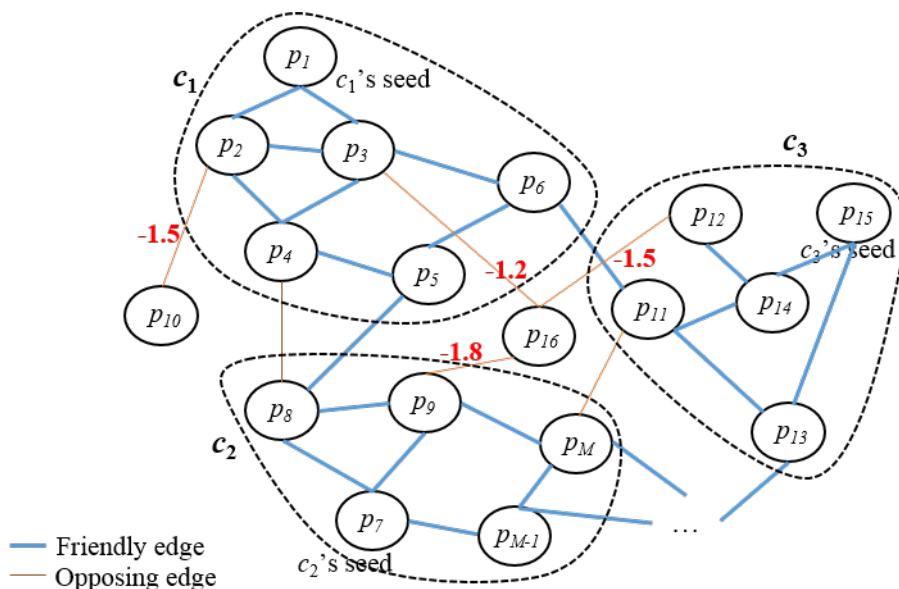


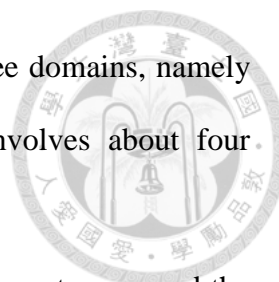
Figure 17. An example of the associations of stance-irrelevant persons

## 4.6 Experimental results of the SCIFNET

In this section, we introduce the data corpus used in the experiments for the SCIFNET; demonstrate the effectiveness of each system component; and compare the SCIFNET's performance with those of other well-known community detection methods and clustering algorithms. Then, we present a stance identification result and discuss the stance-irrelevant persons detected by the SCIFNET.

### 4.6.1 Dataset

As mentioned earlier, topic person stance identification is a relatively new research area, and there is no official corpus for the subject; hence, we compiled a data corpus for evaluations. The corpus comprises thirty topics and 4,996 topic documents, all



downloaded from the Google News. The collected topics cover three domains, namely sport, business issues, and political elections; and each topic involves about four competing stances, as shown in Table 5.

To extract important topic persons mentioned in the topic documents, we used the well-known Stanford Name Entity Recognizer, which tags the person names in an input text. The recognizer extracted 6,648 unique person names for all the topics. We found that a large number of the person names rarely appeared in the topic documents; and the frequency distribution followed Zipf’s law (Zipf, 1949). In other words, there were very few frequent person names. Moreover, as there is no perfect named entity recognizer, several of the infrequent person names were incorrect or ambiguous (e.g., a string intermixed with the name of an organization and the name of a person). To assess our method’s performance accurately, for each of the evaluated topics, we manually removed the false person name entities and only evaluated the first frequent person names whose accumulated frequency reached  $\lambda = 50, 60,$  and  $70$  percent of the total frequency of all the extracted person names. The average number of evaluated person names under each setting of  $\lambda$  is shown in Table 5. All the names represent important topic persons<sup>5</sup>.

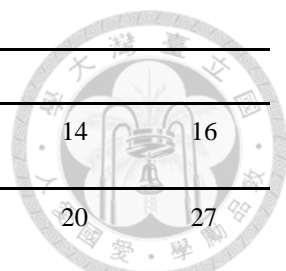
Table 5. The data corpus for the SCIFNET

ID	Topic Title (Date)	# of documents	# of extracted person names	# of evaluated person names for $\lambda$ cumulative frequency		
				$\lambda=50\%$	$\lambda=60\%$	$\lambda=70\%$
$T_1$	The 2011 NFL Conference Finals (2012/01/16-2012/01/24)	104	217	22	30	45
$T_2$	The 2008 NBA Conference Finals (2008/05/20-2008/05/31)	119	93	8	12	15
$T_3$	The 2009 NBA Conference Finals	87	99	9	12	16

<sup>5</sup> <http://weal.im.ntu.edu.tw/SCIFNET.html>

	(2009/05/19-2009/05/30)					
$T_4$	The 2010 NBA Conference Finals (2010/05/16-2010/05/30)	166	162	12	17	20
$T_5$	The 2011 NBA Conference Finals (2011/05/14-2011/05/27)	292	135	9	13	19
$T_6$	The 2012 NBA Conference Finals (2012/05/26-2012/06/10)	233	139	10	13	17
$T_7$	The 2011 MLB Conference Finals (2011/10/7-2011/10/17)	137	173	24	32	42
$T_8$	The 2012 UEFA Champions League (2012/4/24-2012/4/26)	188	144	17	20	27
$T_9$	The 2014 NHL Conference Finals (2014/5/16-2014/6/2)	106	319	17	25	39
$T_{10}$	The 2014 FIFA World Cup semi-finals (2014/7/8-2014/7/10)	380	792	32	49	69
$T_{11}$	IMF meeting to select a new president (2011/05/27-2011/06/05)	150	66	5	11	14
$T_{12}$	2011 OPEC meeting to set oil production quotas (2011/06/06-2011/06/10)	118	167	22	31	43
$T_{13}$	2012 Greek Bailout (2012/11/04-2012/11/09)	69	87	7	11	23
$T_{14}$	Microsoft and i4i lawsuit over patent violation (2011/06/09-2011/06/16)	92	32	8	12	12
$T_{15}$	Banco Espírito Santo Bailout (2014/8/3 – 2014/8/5)	178	239	15	20	27
$T_{16}$	Fox withdraws bid for Time Warner (2014/8/4 – 2014/8/7)	311	372	31	37	44
$T_{17}$	Strike of the Market Basket (2014/7/26 – 2014/8/1)	170	247	12	21	37
$T_{18}$	Amazon/Hachette Fight (2014/8/6 – 2014/8/12)	261	265	17	29	44
$T_{19}$	NCAA Antitrust Lawsuit (2014/8/8 – 2014/8/12)	122	256	15	20	27
$T_{20}$	Fyffes faces rival bid in Chiquita merger deal (2014/8/11 – 2014/8/12)	142	152	11	14	18
$T_{21}$	2012 Russian Presidential Election (2012/02/20-2012/03/06)	94	112	12	17	22
$T_{22}$	2012 French Presidential Election (2012/04/17-2012/04/25)	230	201	17	20	25
$T_{23}$	2012 Mexican Presidential	105	115	10	16	23





	Election (2012/06/29-2012/07/09)					
$T_{24}$	2012 Korean Presidential Election (2012/08/16-2012/08/28)	74	73	12	14	16
$T_{25}$	2014 Afghanistan Presidential Election (2014/6/25 – 2014/7/13)	401	593	12	20	27
$T_{26}$	2014 Indonesian Presidential Election (2014/7/18 -2014/7/23)	173	300	6	14	26
$T_{27}$	2014 Turkish Presidential Election (2014/8/7 – 2014/8/11)	93	151	13	19	26
$T_{28}$	2014 Gaza Strip Crisis (2014/07/20-2014/07/23)	118	431	28	35	46
$T_{29}$	2014 Iraq Crisis (2014/8/1-2014/8/6)	124	297	32	43	61
$T_{30}$	China maritime territorial Dispute (2014/8/10 – 2014/8/13)	159	219	14	20	27

We asked experts to group the evaluated topic persons into stance-coherent clusters and establish a reliable ground truth for the performance evaluation. The kappa statistic which assesses the agreement between the experts is 74.73% and is good enough to conduct reliable evaluations. For the performance evaluation, we used the rand index (Rand, 1971), a popular clustering evaluation metric, because the stance identification method groups topic persons into stance-coherent clusters. There are 1,108,234 person pairs in the dataset. The rand index measures the percentage of all person pairs that are clustered correctly (i.e., if two persons with the same stance are placed in the same cluster or two persons with different stances are placed in different clusters). The higher the score of the rand index, the better the stance identification performance. Because the stance community expansion algorithm depends on seed initialization, we randomly initialize our method, referred to as the SCIFNET, twenty times. The rand index values of all the evaluated topics over the initializations are averaged to obtain the overall stance identification performance. For stance-irrelevant persons detected by the method, we measure their correctness in terms of the F1 score (Manning et al., 2008), which is

the harmonic mean of the detection precision and the detection recall. The score is widely used to evaluate the overall effectiveness of a detection system.



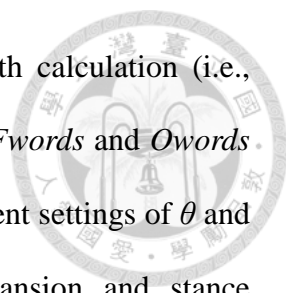
## 4.6.2 System component analysis

### 4.6.2.1 Friendship orientation threshold

Table 6. The lists of *Fwords* and *Owords*

Domain	Stance-friendly word - <i>Fwords</i>	Stance-opposing word - <i>Owords</i>
Business issues	support	criticize
	member	rival
	push	damage
	agreement	rape
	help	fight
	share	campaign
	approve	abuse
	benefit	strike
	partner	reject
	consensus	defend
Political elections	cooperate	campaign
	support	opposite
	help	rival
	member	fraud
	good	accusation
	team	contest
	work	lost
	partner	beat
	advocate	debate
	friend	defeat
Sports	teammate	win
	like	lose
	lead	beat
	best	defend
	good	against
	need	finish
	great	end
	help	guard
	together	defense
	offend	hit

First, we consider the parameter  $\theta$ , which is the threshold of friendship orientation used to establish the edges in a friendship network. In this experiment,  $\theta$  is set between 0.1 and 0.9, and increased in increments of 0.1. Table 6 shows the lists of *Fwords* and *Owords* compiled by two linguistic experts. The stance word lists are used by the stance-oriented correlation coefficient (i.e., Eq. (13)) to compute the stance weight of a



topic document. The parameter  $\beta^{\wedge}$ , used by the friendship strength calculation (i.e., Definition 6), is set at 1. We discuss  $\beta^{\wedge}$  and examine the effects of  $Fwords$  and  $Owords$  later. Figures 18, 19, and 20 show the rand index scores under different settings of  $\theta$  and  $\lambda$ . For each setting of  $\theta$ , we examine stance community expansion and stance community refinement techniques (denoted as SE+SR) in terms of the rand index. We also compare the performance based on stance community expansion only (denoted as SE), i.e., without stance community refinement.

As shown in the figures, the rand index score decreases as  $\lambda$  increases. A large  $\lambda$  implies that the person stance identification is difficult because the setting would include the infrequent topic persons in the stance identification process. As the construction of a friendship network is based on the occurrence of topic persons in the topic documents, including infrequent persons would reduce the quality of the network and therefore affect the stance identification performance. Basically, the rand index score increases as the value of  $\theta$  increases because a large  $\theta$  filters out insignificant friendships between persons to improve the quality of the friendship network. When  $\theta$  is greater than 0.4, the rand index score drops gradually. Connections cannot be established between nodes when  $\theta$  is large. As a result, the friendship network is too sparse to represent informative associations between persons and the stance identification performance is inferior. It is noteworthy that SE+SR performs better than SE. The result demonstrates that stance community refinement resolves the “early merging” problem in stance community expansion and therefore improves the stance identification performance.

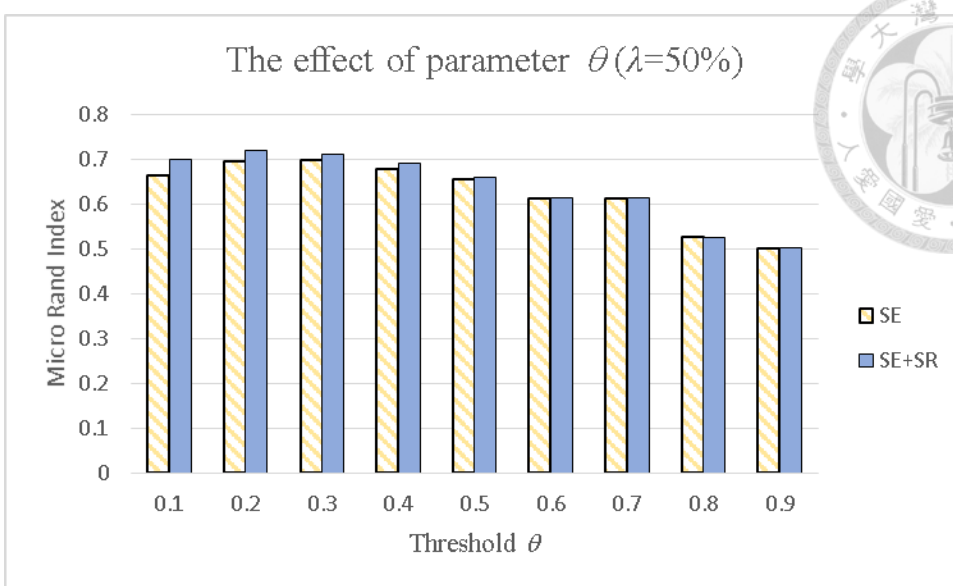


Figure 18. The effect of parameter  $\theta$  under  $\lambda = 50\%$

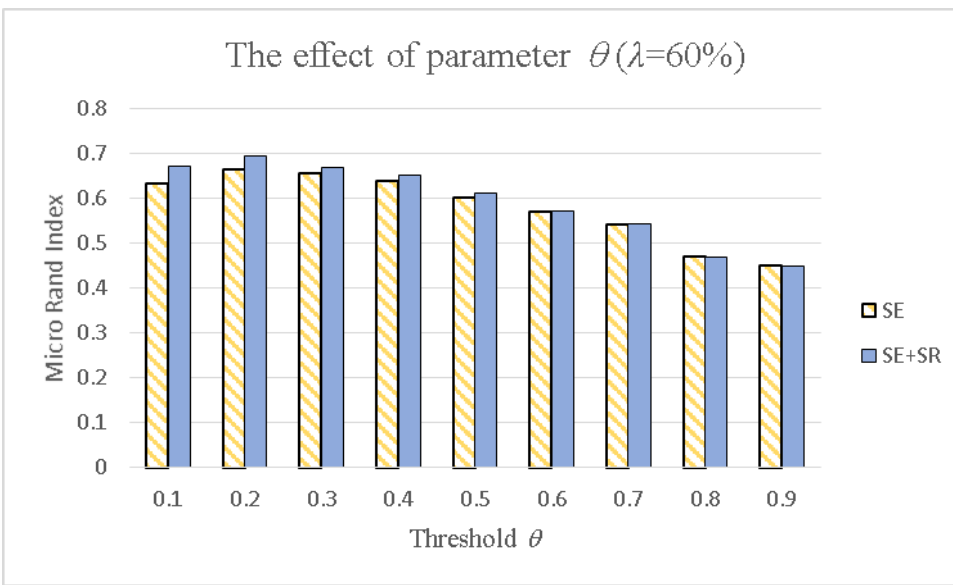


Figure 19. The effect of parameter  $\theta$  under  $\lambda = 60\%$

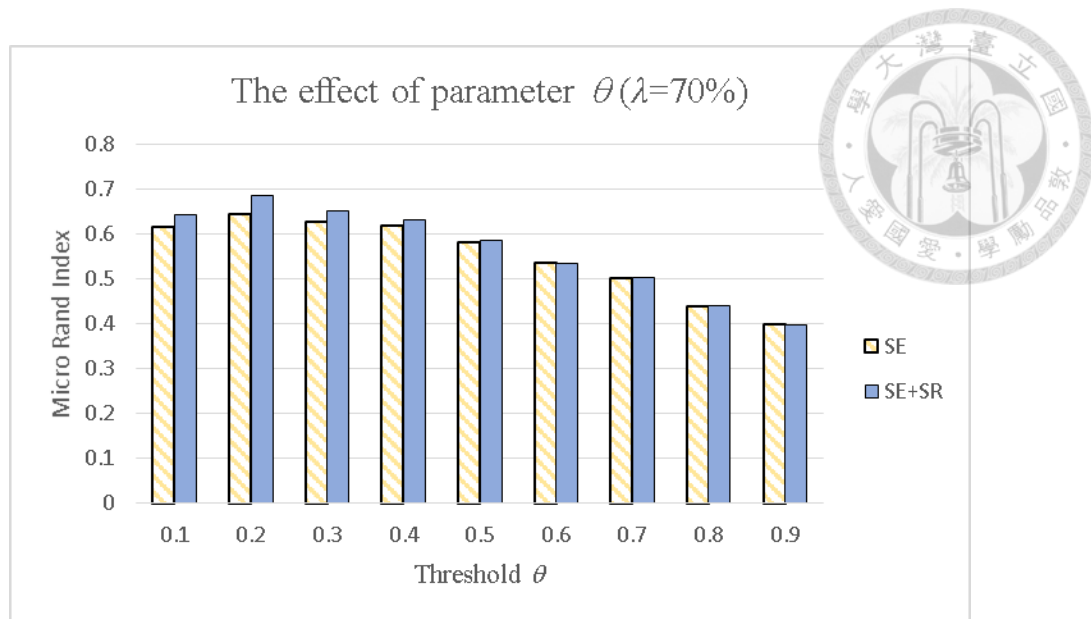


Figure 20. The effect of parameter  $\theta$  under  $\lambda = 70\%$

Figures 21, 22, and 23 show the F1 scores of stance-irrelevant topic person detection under different parameter settings. They also show the corresponding stance-irrelevant person ratio, which is the fraction of topic persons considered stance-irrelevant by our method. Note that the number of stance-irrelevant topic persons detected by SE+SR is the same as that detected by SE. This is because stance community refinement only re-clusters merged boundary nodes, so using it does not affect the stance-irrelevant topic person detection result. For ease of presentation, we only show SE+SR's F1 score and the stance-irrelevant topic person ratio. The F1 scores in the figures are inferior (around 0.2) because the number of stance-irrelevant topic persons in the evaluated topics is small. Hence, a misjudgment of the stance-irrelevant topic persons would reduce the F1 score significantly. The poor F1 scores also indicate that detecting stance-irrelevant topic persons is very difficult. Nevertheless, the scores are still superior to those of many of the community detection methods evaluated in the following experiments. As shown in the figures, a small  $\theta$  value (e.g.,  $\theta = 0.1$ ) always produces a poor F1 score. The reason is that the friendship network constructed by a small  $\theta$  contains many weak friendship edges that cause our method to merge a

stance-irrelevant person with a stance cluster. Increasing the value of  $\theta$  would improve the stance-irrelevant topic person detection performance, but setting it too high (i.e., higher than 0.5) would yield a sparse friendship network. Thus, many important topic persons are incorrectly classified as isolated nodes, which increase the stance-irrelevant topic person ratio. The corresponding F1 score is inferior because most of the detected stance-irrelevant persons are false alarms.

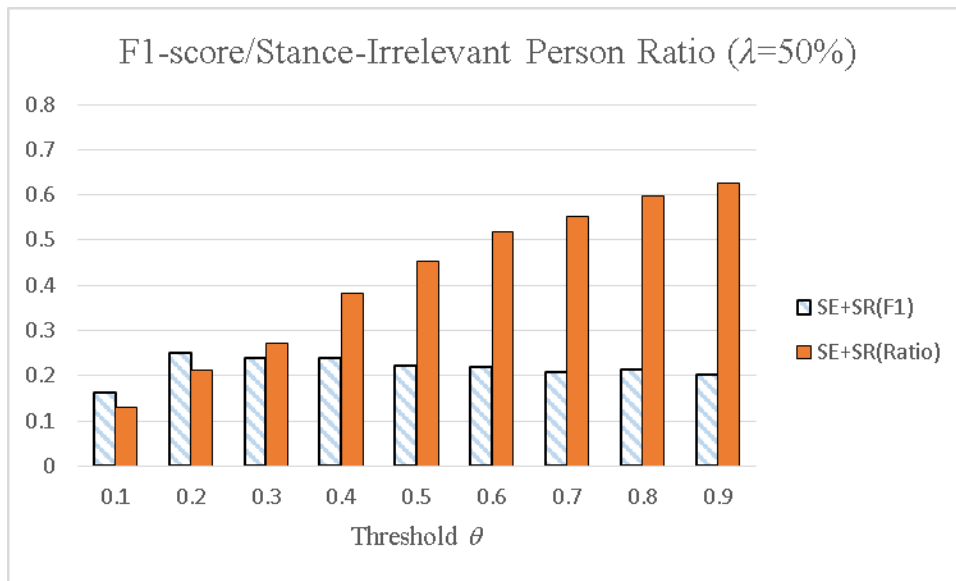


Figure 21. The F1-score/ratio of the detected stance-irrelevant persons under  $\lambda = 50\%$

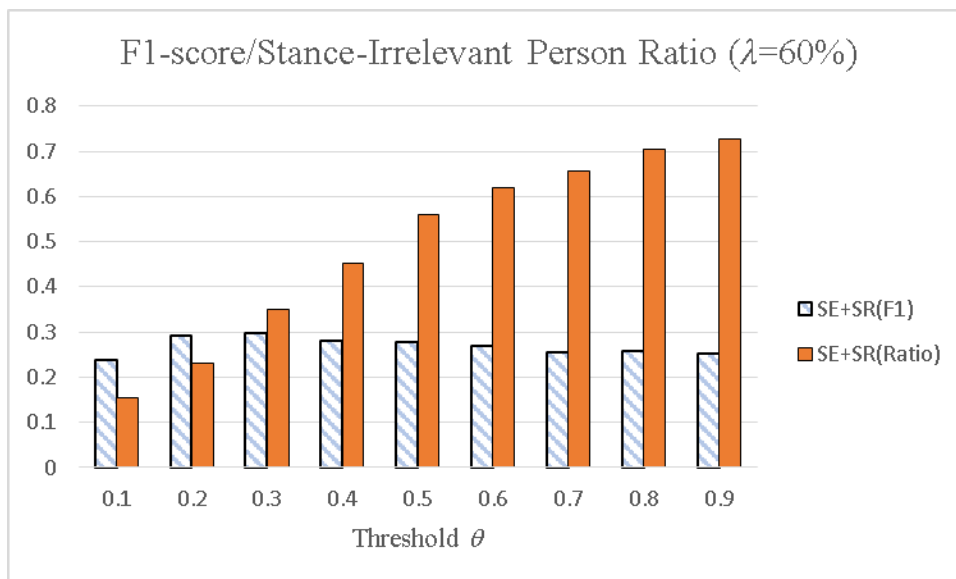


Figure 22. The F1-score/ratio of the detected stance-irrelevant persons under  $\lambda = 60\%$

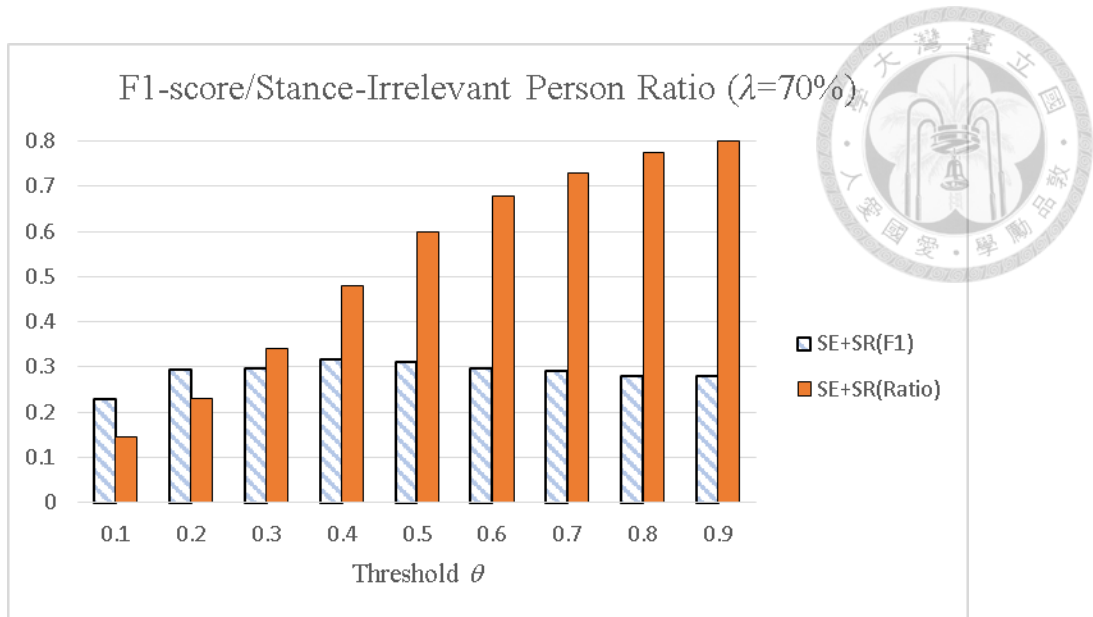


Figure 23. The F1-score/ratio of the detected stance-irrelevant persons under  $\lambda = 70\%$

In summary, a large  $\theta$  increases the ratio of stance-irrelevant topic persons and decreases the rand index score of topic person stance identification. Setting  $\theta$  at 0.2 generally produces good rand index and F1 scores while maintaining a low stance-irrelevant person ratio. Therefore, we set  $\theta$  at 0.2 in the following experiments.

#### 4.6.2.2 Friendship Orientation Threshold using different perspective

In this section, we show the experiments under different threshold for a specific  $\lambda = 70\%$  to discuss the difference between different topic domains: sports, business, and politics.

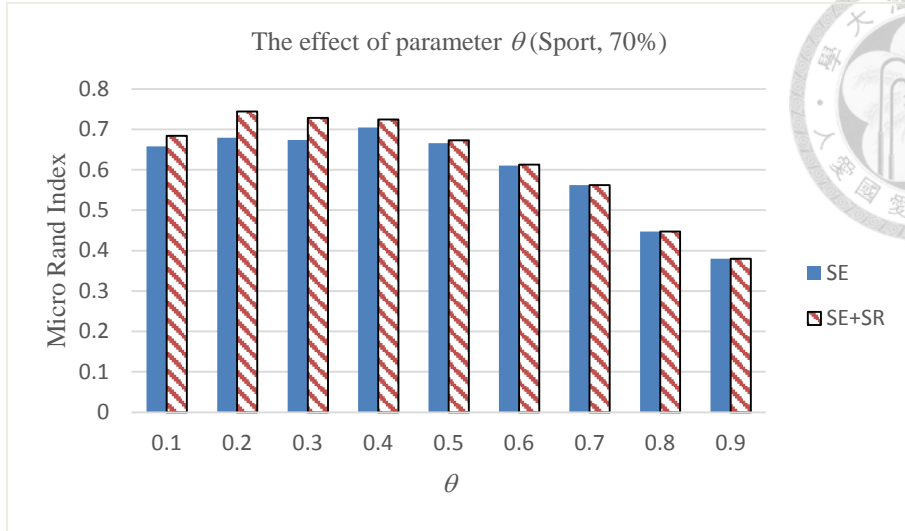


Figure 24. The effect of parameter  $\theta$  on Sports topics under  $\lambda = 70\%$

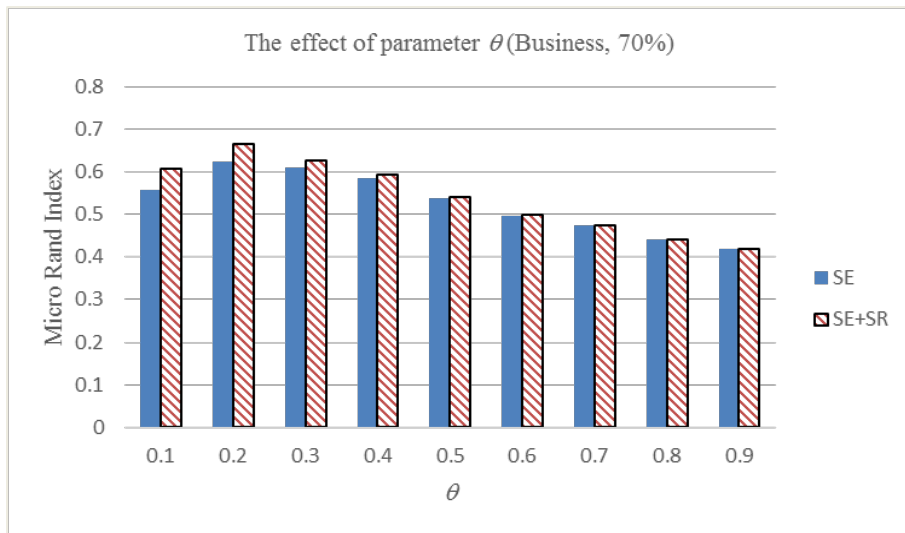


Figure 25. The effect of parameter  $\theta$  on Business topics under  $\lambda = 70\%$

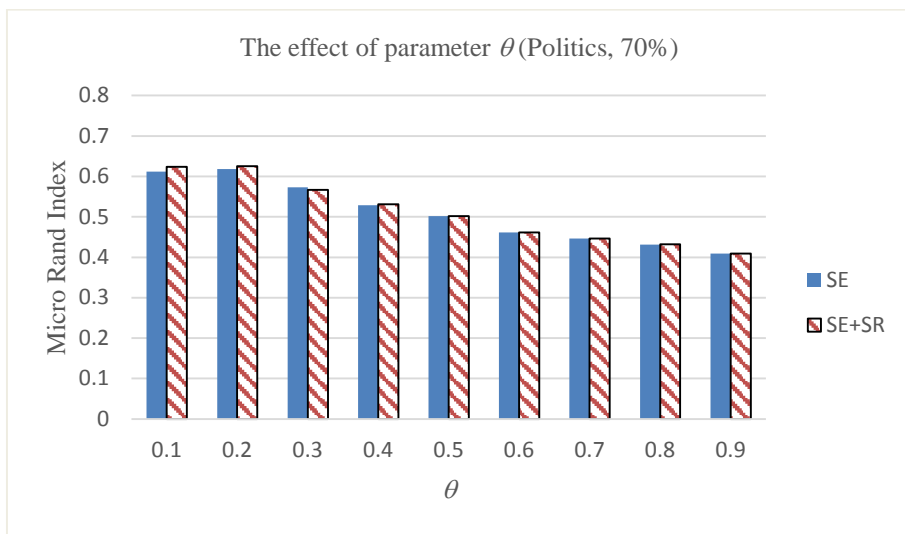


Figure 26. The effect of parameter  $\theta$  on Politics topics under  $\lambda = 70\%$



Obviously, under  $\lambda = 70\%$ , the performance of the Sports topics are superior to the other topics. This is because the Sports news reports the news in terms of the relationships of the teams. For example, when the news mentioned two teams' members, it always contains the description of the competition between the teams, such as how to beat the other team or how to win the game. If the news only report one team, it may discuss the members' situations within the team. When the member mentions the other members, they always praise for their members' performance despite of winning or losing in the last game. Therefore, the performance of the topic person stances in the Sports news is superior to the other topics.

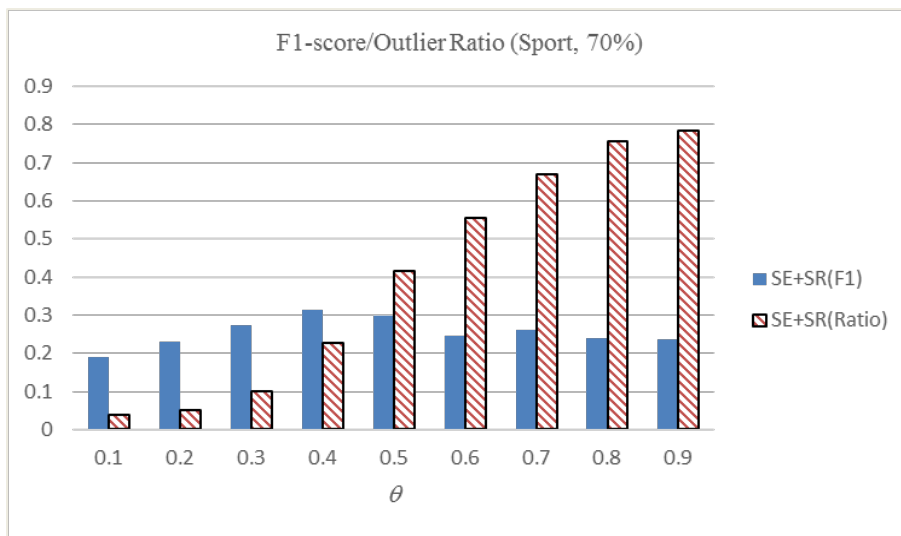


Figure 27. The F1-score/ratio of the detected stance-irrelevant persons on Sports topics under  $\lambda = 70\%$

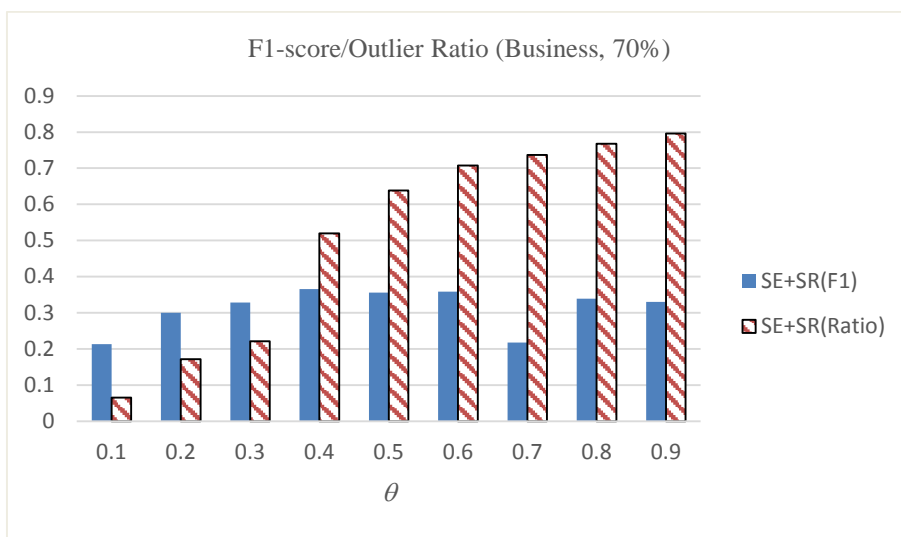


Figure 28. The F1-score/ratio of the detected stance-irrelevant persons on Business topics under  $\lambda = 70\%$

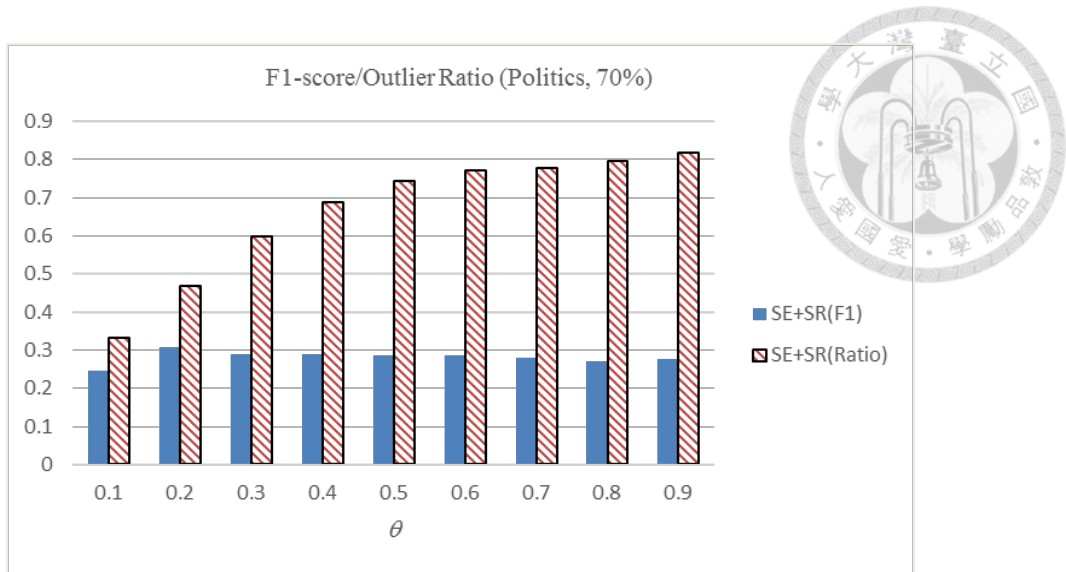


Figure 29. The F1-score/ratio of the detected stance-irrelevant persons on Politics topics under  $\lambda = 70\%$

The trends of the stance-irrelevant detection on different domains under  $\lambda = 70\%$  are similar to Figure 23. Interestingly, the performance of the politics drops dramatically. This is because the political news contains lots of people who advocate their politician, but their friendship strength to the advocated politician may be weak. Hence, when the threshold increases, their connection will be eliminated and the performance drops accordingly.

#### 4.6.2.3 Edge weight evaluation

Next, we discuss the friendship strength (i.e., Definition 6), which combines the friendship orientation and the co-neighboring Jaccard coefficient to compute the weight of a network edge. We evaluate the friendship strength by comparing it with its two constituents. In addition, we assess parameter  $\beta^\wedge$ , which ensures that the friendship strength's exponent factor is not less than 1. As shown in Table 7, the rand index scores under different settings of  $\beta^\wedge$  are very similar. The results imply that the proposed friendship strength is insensitive to the setting of  $\beta^\wedge$ . Nevertheless, setting  $\beta^\wedge$  at 1 usually yields a superior performance, so we use the setting in the following

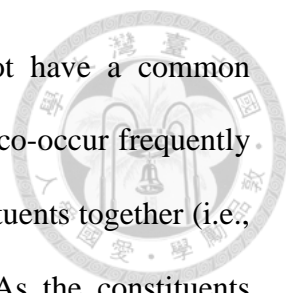
Table 7. Comparison of the edge weighting strategies



$\lambda$	Edge weighting strategy	Rand Index
50%	The friendship orientation	0.709868***
	The co-neighboring Jaccard coefficient	0.669420***
	The friendship strength ( $\beta^{\wedge}=1$ )	<b>0.719063</b>
	The friendship strength ( $\beta^{\wedge}=2$ )	0.717936
	The friendship strength ( $\beta^{\wedge}=3$ )	0.715695*
	The friendship strength ( $\beta^{\wedge}=4$ )	0.713277***
	The friendship strength ( $\beta^{\wedge}=5$ )	0.713057***
60%	The friendship orientation	0.677357***
	The co-neighboring Jaccard coefficient	0.643340***
	The friendship strength ( $\beta^{\wedge}=1$ )	<b>0.695307</b>
	The friendship strength ( $\beta^{\wedge}=2$ )	0.681714***
	The friendship strength ( $\beta^{\wedge}=3$ )	0.683939***
	The friendship strength ( $\beta^{\wedge}=4$ )	0.681764***
	The friendship strength ( $\beta^{\wedge}=5$ )	0.682766***
70%	The friendship orientation	0.654541***
	The co-neighboring Jaccard coefficient	0.627006***
	The friendship strength ( $\beta^{\wedge}=1$ )	<b>0.687692</b>
	The friendship strength ( $\beta^{\wedge}=2$ )	0.663422***
	The friendship strength ( $\beta^{\wedge}=3$ )	0.664411***
	The friendship strength ( $\beta^{\wedge}=4$ )	0.664494***
	The friendship strength ( $\beta^{\wedge}=5$ )	0.664260***

The results marked with \*, \*\* and \*\*\* show, respectively, the improvements in the friendship strength ( $\beta^{\wedge}=1$ ) over the compared strategies with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions(Keller, 2008).

experiments. Surprisingly, the rand index based on the co-neighboring Jaccard coefficient is inferior. This is because the approach tends to underestimate the



association of topic persons. For instance, if two persons do not have a common neighbor, the weight of the edge between them is zero even if they co-occur frequently in the topic documents. It is noteworthy that applying the two constituents together (i.e., the proposed friendship strength) achieves the best performance. As the constituents measure the association between nodes from different perspectives, applying them together identifies the friendship between topic persons accurately and therefore improves the system's performance. For example, in the sports topic "the 2011 NBA Conference Finals," if we simply employ the friendship orientation, the edge weight between Jason Terry and Shawn Marion, who are teammates of Dallas Maverick, would only be 0.280442. By combining the co-neighboring Jaccard coefficient with the friendship orientation, the edge weight increases to 1.448904. The improvement corresponds with the results reported by Jeh and Wisdom (2002) and Antonellis et al. (2008) who demonstrated that the association between nodes is proportional to their co-neighboring level.

#### **4.6.2.4 Stance-oriented correlation coefficient evaluation**

Finally, we evaluate the stance-oriented correlation coefficient (i.e., SOCOR defined in Eq. (13)). The stance-oriented correlation coefficient enhances the traditional correlation coefficient (denoted as COR) by considering a document's stance weight, which is computed by using Turney and Littman's PMI method with the stance words listed in Table 6. Here, we compare our stance-oriented correlation coefficient with the traditional correlation coefficient. Turney and Littman also compiled a semantic orientation word list and used it to determine the semantic orientation of a text unit. To demonstrate the effect of our stance word list, we also compare the system's performance using the semantic orientation word list. In addition, the SentiWordNet is

also a famous dictionary for the sentiment analysis (Esuli et al., 2006; Ohana & Tierney, 2009; Baccianella et al., 2010). We also compare the stance word list with it.

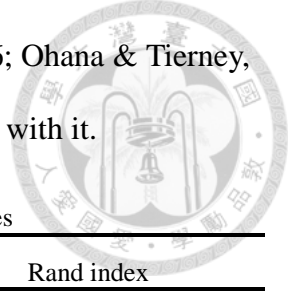


Table 8. Comparison of the correlation coefficient approaches

$\lambda$	The correlation coefficient approach	Rand index
50%	COR	0.703436***
	SOCOR (the stance word list)	<b>0.719063</b>
	SOCOR (the semantic orientation word list)	0.631377***
	SOCOR (SentiWordNet)	0.554219***
60%	COR	0.678924***
	SOCOR (the stance word list)	<b>0.695307</b>
	SOCOR (the semantic orientation word list)	0.632086***
	SOCOR (SentiWordNet)	0.519380***
70%	COR	0.669442***
	SOCOR (the stance word list)	<b>0.687692</b>
	SOCOR (the semantic orientation word list)	0.618314***
	SOCOR (SentiWordNet)	0.503846***

The results marked with \*, \*\*, and \*\*\* show, respectively, the improvements achieved by SOCOR (the stance word list) over the compared approaches with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions.

SOCOR outperforms COR, as shown in Table 8. The results demonstrate that the stances of topic documents are informative for identifying the friendship orientation of topic persons. Notably, SOCOR with the semantic orientation word list and SOCOR with the SentiWordNet are inferior. This is because the lists are used to identify text units that convey positive or negative meanings, and the meanings may not reveal whether the associations between persons are friendly or opposing. For example, in topic  $T_3$ , the document describes the relationships between Lakers' team members and contains the friendly sentence, i.e., *"We found our balance," Gasol said. "We did a good job overall as a group working hard and getting it done. So we'll keep it that way."* The document orientation value is positive when using the list we proposed, but the negative value is obtained by the SentiWordNet.

#### 4.6.2.5 The effect of the adoption all the extracted topic persons

As mentioned above, we evaluated the performance of our method under different  $\lambda$ . However, in this section, we evaluate the  $\lambda$ 's effect when we take the whole extracted topic person names into consideration. The experiments are shown as below.

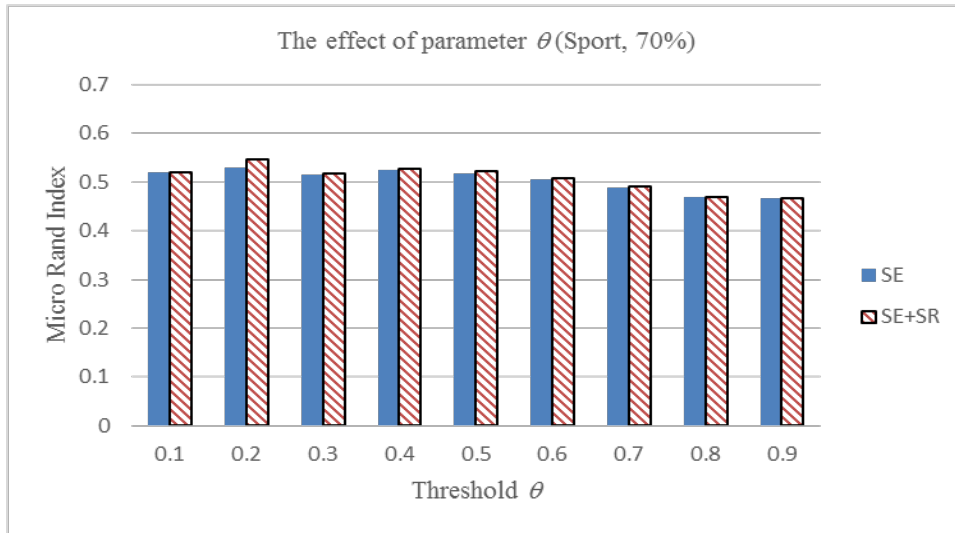
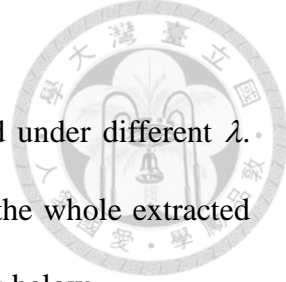


Figure 30. The effect of parameter  $\theta$  includes all the extracted person names on Sports topics under  $\lambda = 70\%$

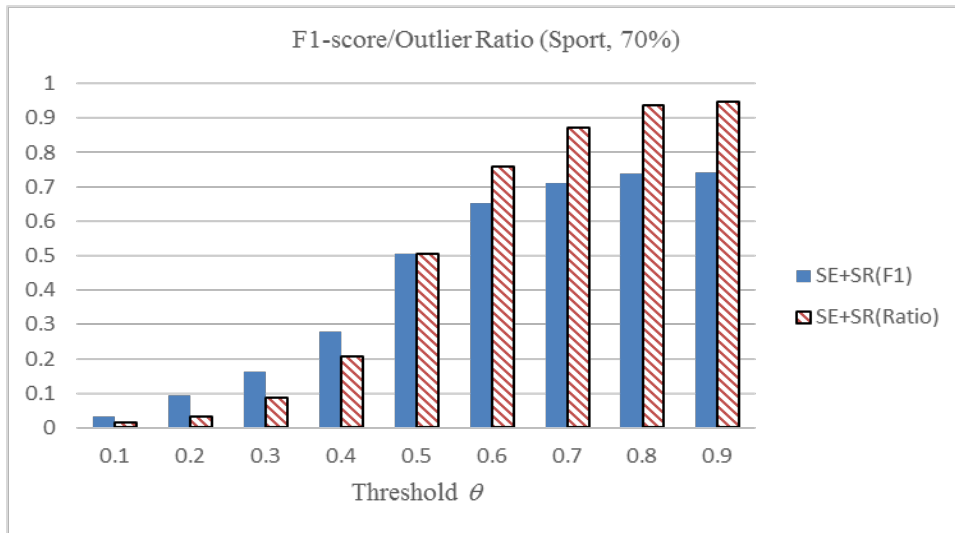
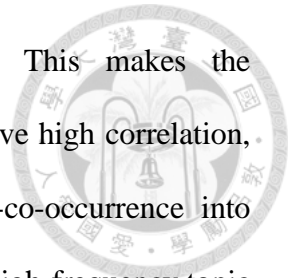


Figure 31. The F1-score/ratio of the detected stance-irrelevant persons includes all the extracted person names on Sports topics under  $\lambda = 70\%$

We found that the low-frequency topic persons make the stance identification of topic persons more difficult because it contains more noisy information. In addition, the

low-frequency topic persons are absent in many documents. This makes the low-frequency topic person and the high-frequency topic persons have high correlation, because the stance-oriented correlation coefficient takes the non-co-occurrence into consideration. As a result, when the threshold increase, the low and high-frequency topic persons will be removed at the same time which makes the performance drop.



#### 4.6.2.6 The effect of the adoption of the other named entities

In this section, we consider more named entities, not only person names, but also organizations and places, to demonstrate the effect of adopting other named entities. We only conduct this experiment on the Sports domain because the Sports domain has the best performance, which can easily reflect the effect of adopting the other named entities. The experiments are shown as below.

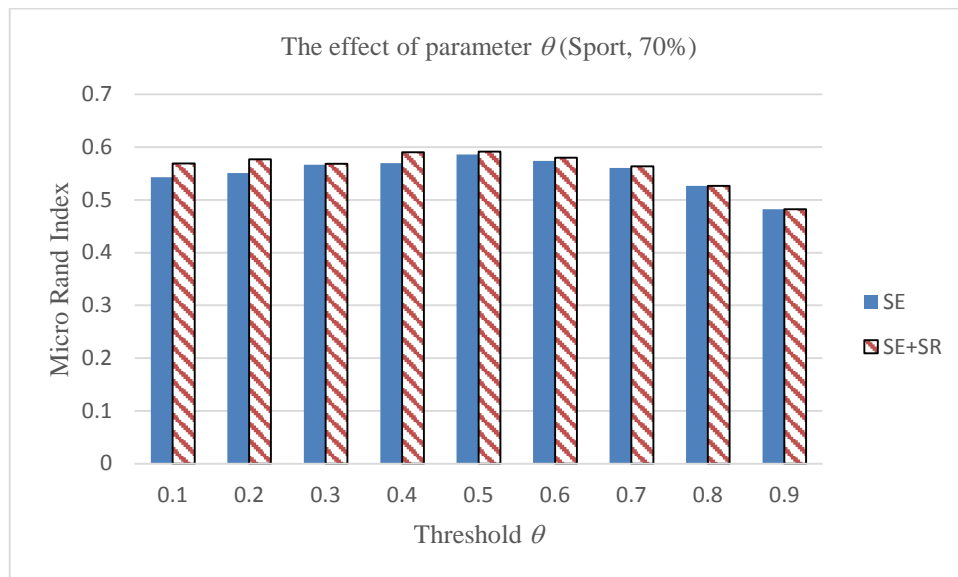


Figure 32. The effect of parameter  $\theta$  includes other named entities on Sports topics under  $\lambda = 70\%$

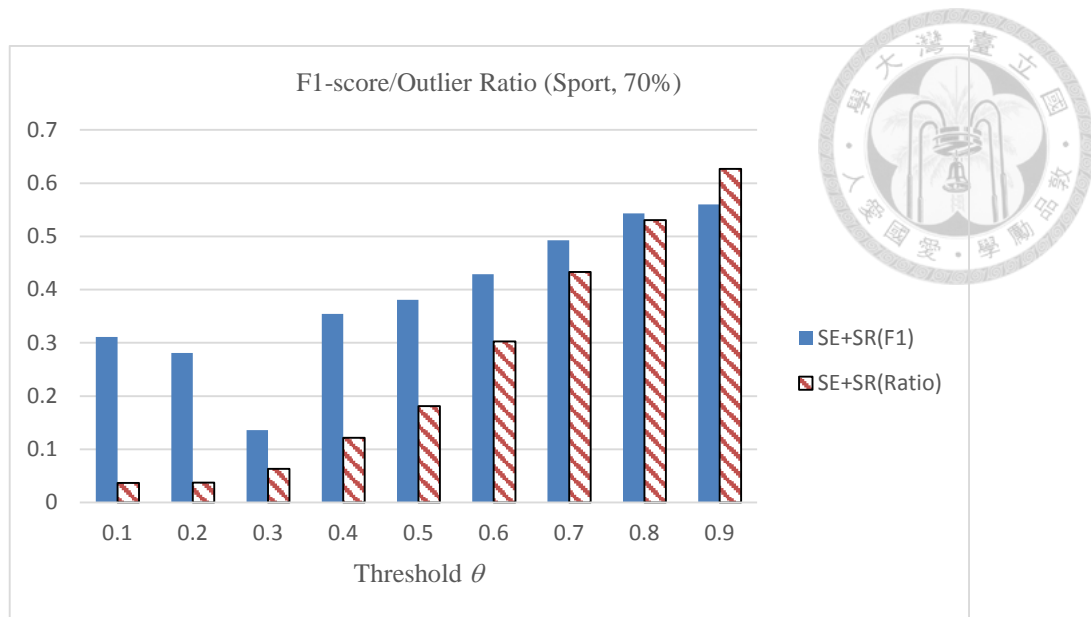


Figure 33. The F1-score/ratio of the detected stance-irrelevant persons includes other named entities on Sports topics under  $\lambda = 70\%$

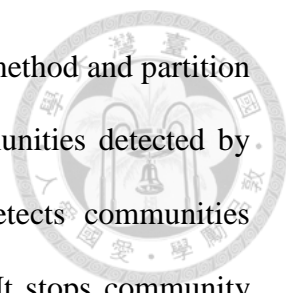
In Figure 32 and 33, we found that the frequencies of the organization names and place names are very high, and make the evaluated person names become fewer, which affects the performance of the topic person stance identification. The higher frequencies of other named entities also make the topic person connection weaker. This is because when the threshold increases, the threshold will filter out the topic persons' connections instead of removing the connection of the organizations or places. It means that the adoption of the other named entities is not helpful for the identification of the topic person stance.

### 4.6.3 Comparison with other methods

#### 4.6.3.1 Stance identification evaluation

In this sub-section, we compare SCIFNET with five well-known community detection approaches: FastModularity (Newman, 2004), SCAN (Xu et al., 2007), CODA (Gao et al., 2010), FEC (Yang et al., 2007), and the signed Modularity (SM) method (Anchuri & Magdon-Ismaïl, 2012). To ensure that the comparisons are fair, all the community





detection methods run on the friendship networks generated by our method and partition each network into  $K$  communities. Note that the number of communities detected by SM sometimes is less than  $K$ . This is because the method detects communities according to the signs of the entries in the principal eigenvector. It stops community detection if the entry signs are all the same. Also note that FEC and SM are designed for signed networks. FastModularity, SCAN, and CODA assume the analyzed networks are unsigned and examine the link structures to detect communities. Our friendship networks contain negative edges. To reduce the influence of negative edges on the methods, we also run the methods on the friendship networks without negative edges. We use the suffix “-neg” to indicate the result without negative edges. For instance, SCAN-neg stands for the result of SCAN on the friendship networks without negative edges. In SCAN, the clustering parameters  $\varepsilon$  and  $u$  are set at 0.5 and 2 respectively, as suggested by (Xu et al., 2007); the link importance parameter of CODA is set at 0.2, as suggested by (Gao et al., 2010); and the parameter  $l$  of FEC is set at 10, as suggested by (Yang et al., 2007).

We also compare two popular clustering algorithms, namely, K-means (Manning et al., 2008) and HAC (Mitchell, 1997). Both algorithms represent a topic person as an  $N$ -dimensional frequency vector in which an entry indicates the frequency that a topic person occurs in a topic document. To measure the association of topic persons, we utilize the cosine similarity (Manning et al., 2008) which is frequently used to determine the similarity of frequency vectors. For HAC, we consider four well-known cluster similarity strategies, namely, single-link, complete-link, average-link, and centroid-link strategies. In addition to the above methods, we compare another baseline method that clusters topic persons randomly. As the clustering results of CODA and K-means depend on their initializations, we randomize both methods twenty times and select the

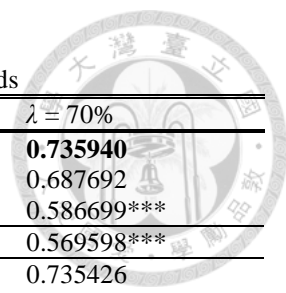


Table 9. The rand index performance of the compared methods

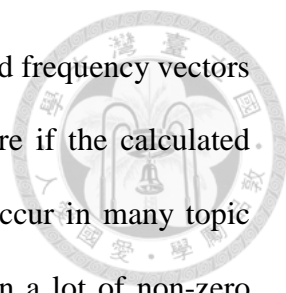
Method	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
SCIFNET (Best)	<b>0.788345</b>	<b>0.750465</b>	<b>0.735940</b>
SCIFNET (Avg.)	0.719063	0.695307	0.687692
SCIFNET (Worst)	0.618197***	0.611962***	0.586699***
Eigen-based Method	0.586826***	0.582201***	0.569598***
Model-based EM(Best)	0.784888	0.749994	0.735426
Model-based EM(Avg.)	0.709152***	0.690678	0.683138**
Model-based EM(Worst)	0.610651***	0.615454***	0.604345***
FastModularity	0.624519***	0.644155***	0.622145***
FastModularity-neg	0.593376***	0.620032***	0.597107***
SCAN	0.622753***	0.652274***	0.679756**
SCAN-neg	0.631180***	0.660077***	0.686941
CODA (Best)	0.720176	0.690943	0.673888***
CODA (Avg.)	0.658576***	0.647904***	0.638469***
CODA (Worst)	0.599230***	0.605666***	0.610629***
CODA-neg (Best)	0.724024	0.708113	0.683023**
CODA-neg (Avg.)	0.658659***	0.660517***	0.653193***
CODA-neg (Worst)	0.605827***	0.619545***	0.624368***
FEC	0.681454***	0.679715***	0.638000***
SM	0.703408**	0.695051	0.686458*
HAC (Single-Link)	0.596756***	0.532265***	0.454518***
HAC (Complete-Link)	0.691589***	0.674059***	0.613611***
HAC (Average-Link)	0.697911***	0.671055***	0.677368**
HAC (Centroid-Link)	0.653381***	0.614823***	0.574120***
K-means (Best)	0.776800	0.749463	0.734419
K-means (Avg.)	0.688634***	0.674775***	0.680654**
K-means (Worst)	0.540682***	0.552726***	0.576607***
Baseline (Avg.)	0.399890***	0.346559***	0.309360***

The results marked with \*, \*\*, and \*\*\* show, respectively, the improvements achieved by SCIFNET (Avg.) over the compared methods with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions.

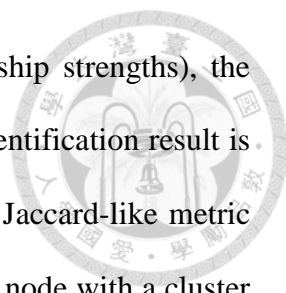
best, worst, and average results for comparison.

We also compare the SCIFNET with our previous work, model-based EM method. Furthermore, for testing the effectiveness of the eigen-based method, we implement a simple method which employs the friendship strength in Definition 6 to construct the friendship matrix, and uses its eigenvector associated with the largest eigenvalue to partition the topic persons into two groups. The procedure will stop until the number of groups reaches the predefined cluster number.

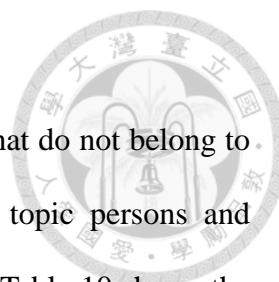
Table 9 shows the comparison results. All the compared methods perform better than the baseline, and our method achieves the best stance identification performance. We observe that HAC and K-means tend to cluster popular topic persons together. This



is because the cosine similarity is the inner product of two normalized frequency vectors (Manning et al., 2008), and it tends to yield a high similarity score if the calculated vectors contain many non-zero entries. As popular topic persons occur in many topic documents, the corresponding normalized frequency vectors contain a lot of non-zero entries. The clustering methods therefore overestimate the association of popular topic persons and group popular, but stance-different, persons together, which degrades the methods' performance. The inferior performance of HAC's single-link strategy is caused by the above defect because the strategy calculates the similarity of two clusters by examining the most similar person pair in the clusters. As a result, the strategy merges clusters containing popular persons even if the clusters represent different stances. By contrast, our method measures the association of topic persons in terms of the stance-oriented correlation coefficient and the co-neighboring strength. Unlike the cosine similarity, the stance-oriented correlation coefficient considers how the occurrences of two topic persons vary jointly in a set of topic documents. Hence, it measures the association of popular topic persons correctly. For instance, in the political topic "the 2012 Korean presidential election," the friendship strength of Park Geun Hye and Park Jie-won, who represented different parties in the election, is -2.11474, but their cosine similarity is 0.984483. It is noteworthy that FastModularity, SCAN, and CODA perform better when negative edges are removed from the friendship networks. As the methods are designed for unsigned networks, negative edges would distract their detection results. The FastModularity algorithm merges nodes into clusters in terms of the modularity measure, which tends to merge clusters that are connected by a lot of edges. However, the measure ignores the edge weights of nodes. Many of the connected edges have small weights that impact the merged cluster's coherency and degrade the algorithm's performance. Our method merges clusters in terms of the merging score (i.e.,



Eq. (15)). As the score is based on the edge weights (i.e., friendship strengths), the nodes in a cluster are highly associated. Consequently, the stance identification result is better than that of the FastModularity algorithm. SCAN employs a Jaccard-like metric to measure the co-neighboring strength between nodes and merges a node with a cluster if their co-neighboring strength is large. Similar to FastModularity, SCAN ignores edge weights, which degrades its performance. In addition to the co-neighboring strength, our friendship strength considers the co-occurrence of nodes in topic documents. SCIFNET therefore outperforms SCAN significantly. While CODA integrates edge weights into its clustering objective function, the weights are based on the cosine similarity of the frequency vectors. Moreover, the objective function simply maximizes the sum of the edge weights in each cluster and ignores the association between the clusters. As a result, CODA groups a lot of popular, but stance-different topic persons, together. In addition to maximizing the association of nodes within clusters, our objective function minimizes the association between clusters. Therefore, SCIFNET achieves a superior stance identification performance. We found that the SM method sometimes cannot produce  $K$  stances (communities) for an evaluated topic because the signs of the entries in the principal eigenvectors are all positive. The method thereby groups persons with different stances together. Besides, the method is based on the modularity which ignores the edge weights. Our method therefore outperforms the SM method. For the comparison with our model-based EM method and the eigen-based method, we found that taking the document orientation into consideration is very effectiveness for identifying the stance of the topic person. The eigen-based method may partition the persons together with the different stances, so as the SCIFNET. However, the SCIFNET can refine the partition results with the stance community refinement to adjust the performance of topic person stance identification.



#### 4.6.3.2 Stance-irrelevant topic person detection evaluation

One function of SCAN and CODA is to detect outliers (i.e., nodes that do not belong to any community). Here, we treat the outliers as stance-irrelevant topic persons and compare their stance-irrelevant topic person detection performance. Table 10 shows the comparison results. Note that CODA uses a clustering objective function to rank the nodes in a network and the last  $\gamma\%$  nodes are denoted as outliers. To ensure a fair comparison, we adjusted  $\gamma\%$  so that the number of stance-irrelevant topic persons detected by CODA is the same as that detected by our method.

As shown in Table 10, the F1 scores of the compared methods are all inferior because we select frequent topic persons for evaluation. All of them are important and influential in the evaluated topics, so very few of them are stance-irrelevant. Consequently, a misjudgment of the stance-irrelevant topic persons would reduce the F1 score dramatically. The inferior performance of the compared methods shows that the detection of stance-irrelevant topic persons is difficult and requires further investigation. Contrary to expectations, SCAN's F1 score is higher than our average F1 score. This is because of SCAN's high detection recall rate. As SCAN clusters nodes in terms of their co-neighboring strength, many weakly-connected nodes are treated as outliers. Consequently, its detection recall is high, which benefits its F1 performance. Nevertheless, our best F1 score is still the best stance-irrelevant topic person detection performance.

Table 10. The F1 performance of stance-irrelevant topic person detection

Method	$\lambda = 50\%$	$\lambda = 60\%$	$\lambda = 70\%$
SCIFNET (Best)	<b>0.358335</b>	<b>0.373005</b>	<b>0.363269</b>
SCIFNET (Avg.)	0.250637	0.292517	0.293951
SCIFNET (Worst)	0.037736	0.102941	0.178218
SCAN-neg	0.259259	0.287356	0.298182
CODA-neg (Best)	0.288889	0.325301	0.316667
CODA-neg (Avg.)	0.248889	0.247590*	0.247083***
CODA-neg (Worst)	0.177778*	0.168675**	0.183333***

The results marked with \*, \*\*, and \*\*\* show, respectively, the improvements achieved by SCIFNET (Avg.) over the compared methods with 90%, 95% and 99% confidence levels based on the Z-statistic for two proportions.

#### 4.6.4 An example of topic person stance identification

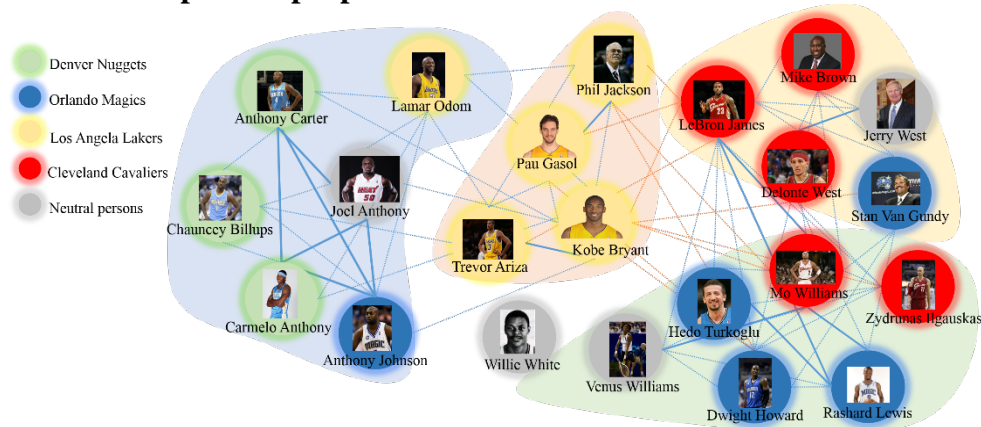
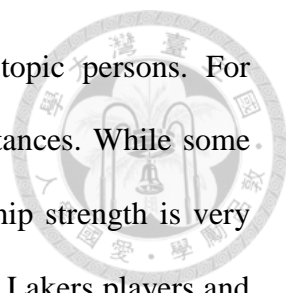


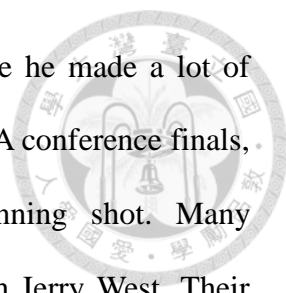
Figure 34. The stance identification result of the 2009 NBA Conference Final ( $\lambda = 70\%$ )

The above experiments quantitatively evaluate the performance of SCIFNET. In this section, we consider a sports topic, namely the 2009 NBA Conference Finals, to assess our stance identification result. The topic covers four basketball teams that competed for the title and we consider each team as a topic stance. Figure 34 shows the constructed friendship network. Stance-irrelevant topic persons are highlighted in gray; and teammates are highlighted in the same color. The blue edges and the orange edges depict friendly associations and opposing associations respectively. Their thickness indicates the friendship strength (i.e., edge weight). As shown in the figure, the



friendship network accurately describes the associations of the topic persons. For instance, the orange edges always connect persons with different stances. While some stance-different persons are connected by blue edges, their friendship strength is very weak. It is noteworthy that many orange edges connect Los Angeles Lakers players and Orlando Magic players. This is because the two teams reached the finals. A large number of the topic documents report the teams' matchup and most of them contain stance-opposing words. As our method utilizes the stance weight of topic documents to measure the friendship strength of topic persons, the matchup-related documents help to capture the opposing orientations of the players.

The colored zones in the figure represent our stance identification results. In this example, the rand index score is 0.762, which show that many topic persons are grouped into stance clusters correctly. Moreover, one topic person (i.e., Willie White) is accurately classified as stance-irrelevant. Notably, our method prevents the teams' franchise players (i.e., Kobe Bryant, Carmelo Anthony, LeBron James, and Dwight Howard), who are also popular topic persons, from being merged. The outcome corresponds with the comparison result presented in the previous section, i.e., the proposed stance-oriented correlation coefficient is effective for measuring the friendship orientation of popular topic persons. We observed that incorrectly-clustered persons often appeared in a few topic documents. For instance, Cleveland player Zydrunas Ilgauskas, who only appeared in 12 topic documents, was clustered as a member of Orlando Magic. We analyzed the phenomenon and found that the stance-oriented correlation coefficient tends to overestimate the friendship of infrequent topic persons. This is because the coefficient is based on the occurrence pattern of topic persons. As infrequent persons are jointly absent from many topic documents, their friendships are overestimated. It is remarkable that Jerry West, an ex-Lakers player, is clustered as a



member of Cavaliers. Jerry West was named “Mr. Clutch” because he made a lot of game-winning shots during his playing career. In Game 2 of the NBA conference finals, Cavaliers player LeBron James made an incredible game-winning shot. Many documents reported the event and tried to place him on a par with Jerry West. Their names thus co-occur frequently in the topic documents so they are clustered together. Interestingly, Venus Williams, a famous tennis player, is included in the experiment. During the matchup of Orlando Magic and Cleveland Cavaliers, Venus Williams was playing in the 2009 French Open. We observed that several topic documents collected from Google News were sports recaps that covered the NBA conference finals as well as the results of the tennis tournament. Consequently, Venus Williams was incorrectly classified as a member of Orlando Magic. The result suggests the analyzed topic documents need to be pure and on-topic. Diverse or noisy documents must be filtered out to enhance the result of topic person stance identification.

#### **4.7 Conclusions of the SCIFNET**

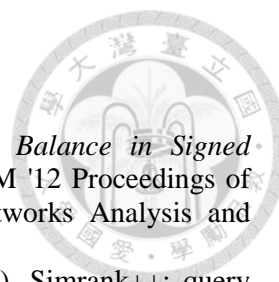
We presented a stance identification method called SCIFNET that constructs a friendship network of topic persons from topic documents automatically. We developed the stance-oriented correlation coefficient to measure the friendship orientation of topic persons. The friendship orientation is then combined with the co-neighboring strength of the topic persons to measure their friendship strengths. Stance community expansion and stance community refinement techniques based on the designed objective function are used to identify stance-coherent clusters of topic persons and identify stance-irrelevant topic persons. The result of experiments on real-world topics demonstrate the effectiveness of SCIFNET and show that it outperforms many well-known community detection and clustering methods.





## 5 Conclusions

The Internet has become a crucial medium for disseminating and acquiring the latest information about topics. However, users are often overwhelmed by the enormous number of topic documents. Basically, times, places, and persons are the key elements of topics. Knowing the associations of topic persons can help readers construct the background knowledge of a topic and comprehend numerous topic documents quickly. In this study, we define the problem of stance identification of topic persons and propose two unsupervised approaches to deal with the problem, namely, model-based EM method and stance identification method based on friendship network. In this study, the number of topic stances is pre-defined. Nevertheless, in our future work, we will incorporate the number of stances into an objective function to determine the appropriate number of stances and stance-group members automatically. We will also consider the context features of topic persons to improve the quality of person stance identification in topics. In the experiment results of the second approach also suggest some interesting areas for future research. For instance, the proposed stance-oriented correlation coefficient is effective in identifying the friendship orientation of popular topic persons; however, it is affected by the frequency sparseness problem of infrequent topic persons. Because infrequent topic persons are jointly absent from a lot of topic documents, the stance-oriented correlation coefficient may overestimate their friendship strength. Reducing the weight of documents when infrequent persons are jointly absent would resolve the overestimation problem.



## 6. References

- Anchuri, Pranay, & Magdon-Ismaïl, Malik. (2012). *Communities and Balance in Signed Networks: A Spectral Approach*. Paper presented at the ASONAM '12 Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining.
- Antonellis, Ioannis, Molina, Hector Garcia, & Chang, Chi Chao. (2008). Simrank++: query rewriting through link analysis of the click graph. *Proceedings of the VLDB Endowment*, 1, 408-421.
- Baccianella, Stefano, Esuli, Andrea, & Sebastiani, Fabrizio, (2010). *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining*. Paper presented at the Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, pp. 2200-2204.
- Barber, David. (2012). *Bayesian Reasoning and Machine Learning*. New York: Cambridge University Press.
- Bartoszynski, Robert, & Niewiadomska-Bugaj, Magdalena. (1996). *Probability and Statistical Inference*. New York, U.S.: John Wiley & Sons.
- Chen, Chien Chin, & Wu, Chen-Yuan. (2010). *Bipolar person name identification of topic documents using principal component analysis*. Paper presented at the Proceeding of the 23rd International Conference on Computational Linguistics.
- Chen, Chien Chin, & Chen, Meng Chang. (2008). *TSCAN: a novel method for topic summarization and content anatomy*. Paper presented at the Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.
- Chen, Chien Chin, & Chen, Meng Chang. (2012). TSCAN: A Content Anatomy Approach to Temporal Topic Summarization. *IEEE Transactions on Knowledge and Data Engineering* 24, 170-183.
- Chen, Chien Chin, Chen, Zhong-Yong, & Wu, Chen-Yuan. (2012). An Unsupervised Approach for Person Name Bipolarization Using Principal Component Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 24, 1963-1976.
- Chen, Jiyang, Zaiane, Osmar R., & Goebel, Randy. (2009a). *Detecting communities in social networks using max-min modularity*. Paper presented at the SIAM International Conference on Data Mining.
- Chen, Jiyang, Zaiane, Osmar R., & Goebel, Randy. (2009b). *Local Community Identification in Social Networks*. Paper presented at the International Conference on Advances in Social Network Analysis and Mining.
- Clauset, Aaron, Newman, M. E. J., & Moore, Christopher. (2004). Finding community structure in very large networks. *Physical Review E*, 70, 066111.
- Ding, C.H.Q., He, Xiaofeng, Zha, Hongyuan, Gu, Ming, & Simon, H.D. (2001). *A min-max cut algorithm for graph partitioning and data clustering* Paper presented at the Proceedings IEEE International Conference on Data Mining.
- Ding, Xiaowen, Liu, Bing, & Yu, Philip S. (2008). *A holistic lexicon-based approach to opinion mining*. Paper presented at the Proceedings of the international conference on Web search and web data mining (WSDM).
- Donath, W. E., & Hoffman, A. J. (1973). Lower Bounds for the Partitioning of Graphs. *IBM Journal of Research and Development*, 17, 420-425.
- Esuli, Andrea, & Sebastiani, Fabrizio, (2006). *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*. Paper presented at the Proceedings of LREC-06, 5<sup>th</sup> Conference on Language Resources and Evaluation, Genova, IT, 2006. European Language Resources Association (ELRA), Paris, FR, pp. 417-422.
- Feng, Ao, & Allan, James. (2007). *Finding and linking incidents in news*. Paper presented at the

- Proceedings of the sixteenth Conference on information and knowledge management.
- Figueiredo, M., & Jain, A. (2002). Unsupervised learning of finite mixture models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24, 381-396.
- Ganapathibhotla, Murthy, & Liu, Bing. (2008). *Mining opinions in comparative sentences*. Paper presented at the Proceedings of the 22nd International Conference on Computational Linguistics
- Gao, Jing, Liang, Feng, Fan, Wei, Wang, Chi, Sun, Yizhou, & Han, Jiawei. (2010). *On community outliers and their efficient detection in information networks*. Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.
- Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99, 7821-7826.
- Godbole, Namrata, Srinivasaiah, Manjunath, & Skiena, Steve. (2007). *Large-Scale Sentiment Analysis for News and Blogs*. Paper presented at the International Conference on Weblogs and Social Media.
- Harris, Zellig. (1954). Distributional structure. *Word*, 10, 146-162.
- Hatzivassiloglou, Vasileios, & McKeown, Kathleen R. (1997). *Predicting the semantic orientation of adjectives*. Paper presented at the Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics
- Hofmann, Thomas. (1999). *Probabilistic latent semantic indexing*. Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval
- Hu, Mingqing, & Liu, Bing. (2004). *Mining Opinion Features in Customer Reviews*. Paper presented at the Proceedings of the 19th national conference on Artificial Intelligence.
- Jeh, Glen, & Widom, Jennifer. (2002). *SimRank: a measure of structural-context similarity*. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.
- Kanayama, Hiroshi, & Nasukawa, Tetsuya. (2006). *Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis*. Paper presented at the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.
- Keller, Gerald. (2008). *Statistics for Management and Economics: Cengage Learning*.
- Kim, Soo-Min, & Hovy, Eduard. (2004). *Determining the sentiment of opinions*. Paper presented at the Proceedings of the 20th international conference on Computational Linguistics
- Ku, Lun-Wei, Liang, Yu-Ting, & Chen, Hsin-Hsi. (2006). *Opinion Extraction, Summarization and Tracking in News and Blog Corpora*. Paper presented at the Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs.
- Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, 39(4), pp. 885-916.
- Liu, Bing. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1), pp. 1-167.
- Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.
- Manning, Christopher D., & Schütze, Hinrich. (1999). *Foundations of Statistical Natural Language Processing*: The MIT Press.
- Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, & Miller, Katherine J. (1990). Introduction to WordNet : An On-line Lexical Database. *International Journal of Lexicography*, 3, 235-244.
- Mitchell, Tom. (1997). *Machine Learning*. Maidenhead: McGraw-Hall.
- Nallapati, Ramesh, Feng, Ao, Peng, Fuchun, & Allan, James. (2004). *Event threading within news topics*. Paper presented at the Proceedings of the thirteenth ACM international conference on Information and knowledge management.
- Newman, M. E. J. (2001). Scientific collaboration networks: I. Network construction and

- fundamental results. *Physical Review E*, 64, 016131.
- Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 066133.
- Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74, 036104.
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.
- Ohana, B., & Tierney, B., (2009). *Sentiment Classification of Reviews Using SentiWordNet*. IT&T Conference.
- Papadopoulos, Symeon, Kompatsiaris, Yiannis, Vakali, Athena, & Spyridonos, Ploutarchos. (2012). Community detection in Social Media. *Data Mining and Knowledge Discovery*, 24, 515-554.
- Pernkopf, F., & Bouchaffra, D. (2005). Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 1344-1348.
- Rand, William M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66, 846-850.
- Shi, Jianbo, & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 888-905.
- Stone, Philip J., Dunphy, Dexter C., Smith, Marshall S., & Ogilvie, Daniel M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*: The MIT Press.
- Traag, V. A., & Bruggeman, Jeroen. (2009). Community detection in networks with positive and negative links. *Physical Review E*, 80, 036115.
- Turney, Peter D., & Littman, Michael L. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21, 315-346.
- White, Scott, & Smyth, Padhraic. (2005). *A Spectral Clustering Approach To Finding Communities in Graphs*. Paper presented at the Proceedings of SIAM International Conference on Data Mining.
- Wu, Chien-Fu Jeff. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11, 95-103.
- Wu, Fa-Yueh. (1982). The Potts model. *Review of Modern Physics*, 54, 235-268.
- Xu, Xiaowei, Yuruk, Nurcan, Feng, Zhidan, & Schweiger, Thomas A. J. (2007). *SCAN: a structural clustering algorithm for networks*. Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.
- Yang, Bo, Cheung, William, & Liu, Jiming. (2007). Community Mining from Signed Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, 19, 1333-1348.
- Yang, Tianbao, Jin, Rong, Chi, Yun, & Zhu, Shenghuo. (2009). *Combining link and content for community detection: a discriminative approach*. Paper presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining.
- Zipf, George Kingsley. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.