Knowledge-Based Systems 37 (2013) 366–377

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A sock puppet detection algorithm on virtual spaces

Zhan Bu *, Zhengyou Xia, Jiandong Wang

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:
Received 5 March 2012
Received in revised form 26 July 2012
Accepted 18 August 2012
Available online 29 August 2012

Keywords:
Virtual spaces
Sock puppet
Authorship identification
Hypothesis test
Community detection

ABSTRACT

On virtual spaces, some individuals use multiple usernames or copycat/forge other users (usually called "sock puppet") to communicate with others. Those sock puppets are fake identities through which members of Internet community praise or create the illusion of support for the product or one's work, pretending to be a different person. A fundamental problem is how to identify these sock puppets.

In this paper, we propose a sock puppet detection algorithm which combines authorship-identification techniques and link analysis. Firstly, we propose an interesting social network model in which links between two IDs are built if they have similar attitudes to most topics that both of them participate in; then, the edges are pruned according a hypothesis test, which consider the impact of their writing styles; finally, the link-based community detection for pruned network is performed. Compared to traditional methods, our approach has three advantages: (1) it conforms to the practical meanings of sock puppet community; (2) it can be applied in online situation; (3) it increases the efficiency of link analysis. In the experimental work, we evaluate our method using real datasets and compared our approach with several previous methods; the results have proved above advantages.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Recently, online virtual networks have gained significant popularity and are now among the most popular sites on the Web [1,2]. Users with similar interests can join the same board, open a topic, and other users comment on this topic so as to continue the discussion. The rapid development and globalization of online virtual networks have improved and enriched our lives of entertainment, but it also facilitates cyber deceptions, such as users' ID theft, article counterfeit, and swindle. Between 2000 and 2003 [32], John Lott, author of More Guns, Less Crime, made numerous posts under the sock puppet name "Mary Rosh". "Rosh" praised Lott's views and disputed with his critics on Usenet, posting laudatory reviews of Lott's books and panning those of his rivals. Lott admitted he had used the name "Mary Rosh" to defend himself but claimed the book reviews were written by his son and wife. In September 2011 [33], Johann Hari, a leading columnist for the British newspaper The Independent, publicly apologized for having used a pseudonym, David Rose, with Wikipedia screen name David r of Meth productions, to add positive material to the Wikipedia article about himself and negative material to Wikipedia articles about people with whom he had disputes. On virtual spaces, individuals using multiple usernames to communicate with others are usually called "sock puppet". Those sock puppets are fake identities through which members of Internet community praise or create the illusion of support for the product or one's work, pretending to be a different person. A fundamental problem is how to identify these sock puppets.

Traditional authorship-identification techniques have demonstrated a high level of discriminatory potential in the authorship-identification of a limited number of articles or texts [3–9,34]. Given a set of writings of different authors, we assign a new piece of writing to one of them. This problem can be considered as a classification problem. The essence of this classification is to identify a set of features that remain relatively constant for a large number of writings created by the same person. Once a feature set has been chosen, a given writing can be represented by an n-dimensional vector, where n is the total number of features. Then, we can apply many analytical techniques to determine the category of a new vector created on the basis of a new piece of writing. Hence, the feature set and the analytical techniques may significantly affect the performance of authorship identification.

Compared to realistic environment, the online situation is more complicated in the sheer amount of cyber users and activities. Massive amounts of the real data collected from online societies are network structured. So, it is not sufficient and reliable to apply traditional authorship-identification techniques across online network society. Community detection, as a major topic in social network analysis, may provide a great inspiration for sock puppet detection. As members of the same Internet community may have common hobbies, social functions, occupations, interests on some topics, viewpoints, etc.; their sock puppets are more likely to appear in the same posts or comment on the same topics. Most

* Corresponding author.
E-mail address: buzhan@nuaa.edu.cn (Z. Bu).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.08.016

previous papers [10–15] on the subject of community detection mainly focus on link analysis or topological structure of the network. The relationship between two IDs is often directly measured according to their interactive times, which confuses the meanings of the sock puppet community.

In this paper, we propose a sock puppet detection algorithm which combines authorship-identification techniques and link analysis. Firstly, we propose an interesting social network model in which links between two IDs are built if they have similar attitude to most topics. Then, the edges are pruned according to a hypothesis testing: (1) Given two IDs who are connected in the network, particular comment sets are extracted from dataset respectively to check their writing features; (2) the null hypothesis is that these two sets are from the same person; (3) the test value T is calculated; (4) in a given test level α, if the null hypothesis is true, we reserve their edge, otherwise, we remove it. Finally, the link-based community detection for pruned network is performed. Compared to the traditional work, our approach can identify similar sock puppets of a certain person, ally or company. Such findings can be used to prevent cyber deceptions, such as users' ID theft, article counterfeit, and swindle.

The remainder of this paper is organized as follows: Section 2 reviews the related work. Problem description is given in Section 3. Section 4 presents the sock puppet detection algorithm based on authorship-identification techniques and link analysis. In order to verify our approach, we conducted extensive experiments on real life data sets. The experimental design and results analysis are given in Section 5. Finally, we conclude the paper in Section 6.

2. Related work

As detection of sock puppets has not been addressed in the current literature, some work can be used as an advising in sock puppet detection, most of which focus on identification techniques and social network analysis. We briefly review the related work as follows:

To deal with a larger number of features pertaining to authorship or article identification models, many analytical approaches have been adopted so far. The two most commonly used techniques are statistical analysis and machine learning approaches. In the nineteenth century, article identification is used to differentiate between the pieces of Shakespeare, Marlowe and Bacon [3,4]. Perhaps the most foundational work in this field is conducted by Mostteller and Wallace [5,6]. They use authorship identification to correctly attribute the twelve disputed Federalist Paper. Recently, this method also has been increasingly applied to online material due to augmented misuse of the Internet. De Vel et al. conduct a series of experiments on authorship identification of emails [7,8]. Their studies provide an important foundation for the application of authorship identification techniques to the internet medium. Zheng et al. expand De Vel et al.'s efforts by adding the multidimensional space in their study of English and Chinese web forum messages [9].

Those works have gained success in some applications but they mainly focus on the identification of a limited number of articles or texts. Compared to realistic environment, the online situation is further complicated by the sheer amount of cyber users and activities. It is not sufficient or reliable to apply traditional authorship identification techniques across online network society. Some researchers focus on finding communities in online social networks [10–15]. As members of the same Internet community may have common hobbies, social functions, occupations, interests on some topics, viewpoints, etc.; their sock puppets are more likely to appear in the same community. A framework for analysis of user activity on an interactive website is proposed by Zeng et al. in [10]. Du et al. propose ComTector to detect the communities efficiently in large-scale social networks [11]. To discover the discussion topics of social networks, McCallum et al. present the Author–Recipient-Topic model [12]. Tian et al. propose OLAP-style aggregation strategies to partition the graph according to attribute similarity, so that nodes within one community share the same attribute values [13]. Zhang et al. propose a topic oriented community detection approach which combines both social objects clustering and link analysis [14].

The above methods aim to group members interested in common topics into one community. However, they confuse the meanings of the sock puppet community. In reality, one's sock puppets may not communicate with each other frequently, but reply to a certain ID together. What is more, those sock puppets should have similar writing styles and consistent views to certain topics. The existing approaches mentioned above consider only one aspect but ignore the other. To address this problem, we propose a sock puppet detection algorithm which combines these two bodies of work.

3. Problem description

A sock puppet is an online identity used for purposes of deception. The term—a reference to the manipulation of a simple hand puppet made from a sock—originally referred to a false identity assumed by a member of an internet community who spoke to, or about himself while pretending to be another person. The term now includes other uses of misleading online identities, such as those created to praise, defend or support a third party or organization. [35] Sock puppets are very prevalent in popular online environment such as blog, BBS and SNS.

Fig. 1a shows a tree structure corresponding to a small thread of depth 4. Labels denote the user who writes the contribution and valid comments are shown within the gray region. The post triggers three responses from users A, C and D. At the second nesting level, eight comments appear. At the third level, there are still seven comments and finally, there is one last comment from C. The attitude of every comment can be represented using + or −, with + denoting a user is supportive to the viewpoint and − otherwise. Fig. 1b is a social network excavated from original thread of comments, including seven nodes and nine edges. The nodes represent the members involved in the social activities and the edges represent the social relations of interactions or communications. The weight attached to each edge represents the strength of connections between the corresponding members. Fig. 1c shows the result of discovered communities based on link analysis. We can see that members within a community are connected, but their opinions are different. In the left community of Fig. 1c, user B is strongly opposed to the viewpoints of user A, in reality, these two IDs may not be from one's self, ally or company.

To sum up, existing studies on community detection have reported the infeasibility of sock puppet detection tasks. In reality, one's sock puppets may not communicate with each other frequently, but reply to a certain ID together. Fig. 1d shows an ideal result. The members within one community have similar writing feature, meanwhile their opinions to a certain topic are basically consistent. This is the result we aim to achieve in this paper. To solve this challenging task, the following questions need first considering:

(*a*) *Network*. As one's sock puppets may not communicate with each other frequently, traditional network models using direct interactive frequency to represent the strength of connections may not be suitable for sock puppet detection. So, a new social network model needs proposing.

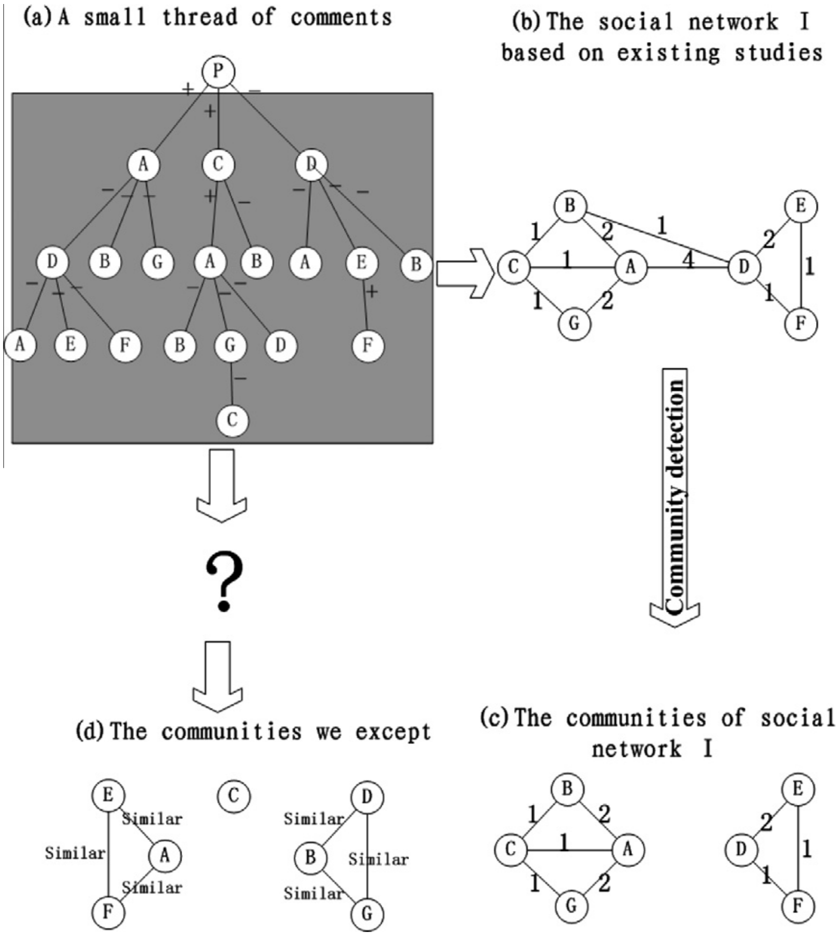

(a) A small thread of comments
(b) The social network I based on existing studies
Community detection
?
(d) The communities we except
(c) The communities of social network I
P
A
C
D
B
G
E
F
Similar
+
−
1
2
4

Fig. 1. An example to illustrate the motivation of our work.

(*b*) *Similarity*. The similarities between two IDs include two meanings: (1) the basic perspectives to a certain topic should be consistent; (2) the writing feature of the two IDs should also be similar.

4. Sock puppet detection approach

We first propose an interesting social network model in which links between two IDs are built if they have similar attitudes to most topics. Next, the edges are pruned according to a hypothesis testing. Finally, the link-based community detection for pruned network is performed to identify similar sock puppets.

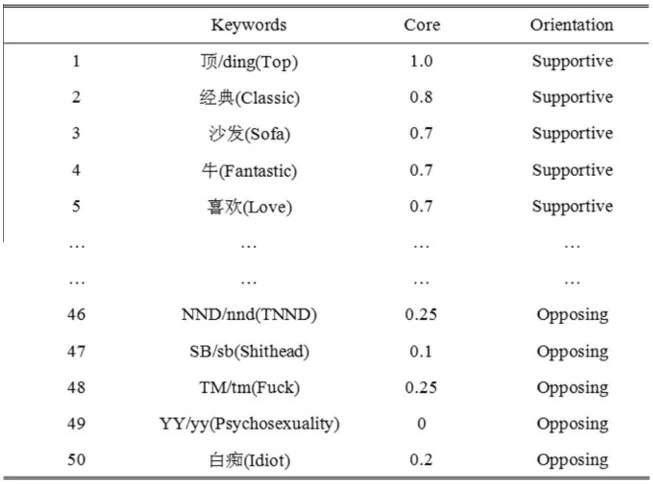
4.1. Similar-view network

In general, a network is built according to the implicit relations between the author of a comment and the user who replies to it. To improve the quality of the resulting network, some of the comments need filtering according to the following criteria:

(1) Those self-replies should also be filtered;
(2) Anonymous comments will not be reserved.

In a social network, every registered user identification (ID) corresponds to a node $i \in V$ in a graph $G = \langle V,E \rangle$. An edge $(i,j) \in E$ represents a social relation between two users that results from their comment activity. Traditional network models use direct interactive frequency to represent the strength of connections between the two IDs; it may not be suitable for sock puppet detection as shown in Fig. 1c. In reality, one's sock puppets may not communicate with each other frequently, but reply to a certain ID together. Moreover, their attitudes to a certain topic should be similar. Our previous research [15] shows that the implicit orientation of every comment can be mostly appraised by several emotional

Table 1
Emotional phrases with English version in parentheses.



	Keywords	Core	Orientation
1	顶/ding(Top)	1.0	Supportive
2	经典(Classic)	0.8	Supportive
3	沙发(Sofa)	0.7	Supportive
4	牛(Fantastic)	0.7	Supportive
5	喜欢(Love)	0.7	Supportive
…	…	…	…
…	…	…	…
46	NND/nnd(TNND)	0.25	Opposing
47	SB/sb(Shithead)	0.1	Opposing
48	TM/tm(Fuck)	0.25	Opposing
49	YY/yy(Psychosexuality)	0	Opposing
50	白痴(Idiot)	0.2	Opposing

words. Those emotional words basically include two types: supportive and opposed. In Table 1, we roughly identify several terms or phrases, with English version in parentheses, from the public discussions as either supportive or opposing. Every term/phrase is assigned with a value between 0 and 1 according to their tone manually. A higher value corresponds to a greater degree of support; if the phrase is neutral, we assigned it a value of 0.5. Thus, every phrase has an associated numerical "trust". For a given comment from one ID to the other, we can determine the implicit orientation by counting the number of positive or negative words in it (if there are several emotional words in one comment, we take the average). Accordingly, the "trust" from user i to a given topic p can be calculated as:

$$trust_i^p = \frac{\sum_{k=1}^{n^p} o_k^p}{n^p} \tag{1}$$

where o_k^p is the implicit orientation of one comment under the topic p, and n^p is reply number from user i to the given topic. For a given ID pair i and j, they may together reply to some topics; we use a topic set $P^{i,j}$ to present them. The attitude consistency of the given ID pair to those topics can be acquired by a simple statistical method as follow:

$$AC_{i,j} = \frac{\sum_{p \in P^{i,j}} \delta\left(trust_i^p, trust_j^p\right)}{n^{P^{i,j}}} \tag{2}$$

where $n^{P^{i,j}}$ is the number of topics which are replied together by user i and user j. And $\delta(x,y)$ is a judgment function determined by x and y, which obeys

$$\delta(x,y) = \begin{cases} 1 & x > 0.5, y > 0.5 \ or \ \mathrm{x} < 0.5, \mathrm{y} < 0.5 \\ 0 & otherwise \end{cases} \tag{3}$$

The range of $AC_{i,j}$ is between 0 and 1, and a higher value corresponds to a greater degree of consistency of the given ID pair to topics. To confuse forum administrators, some sock puppets even pretend that they have different attitudes to some topic. In this paper, if the value of $AC_{i,j}$ is greater than 0.5, the attitudes of user i and user j are consistent. As shown in Fig. 2a, user i and user j have same attitudes to topic A, but have different attitudes to topic B. The attitude consistency of the given ID pair $AC_{i,j}$ is 0.5, which is not greater than 0.5; therefore, we think their attitudes are inconsistent. Some other examples can be seen in Fig. 2b–d.

Let n_{ij} be the number of times that user i writes a comment to user j, and $a_{ij} \in \{+,-\}$ be its orientation. An undirected edge exists between users i and j if (a) $n_{ip} > 0$ and $n_{jp} > 0$ where $i \neq p$, $j \neq p$; (b) $AC_{i,j} > 0.5$; (c) $n_{ij} + n_{ji} < \phi$. Where $n_{ij} + n_{ji}$ is the interactive time between user i and user j. We use a threshold ϕ to restrain this value in a given range, which means that the connected users do not communicate with each other frequently. The network established according to this way is called similar-view network (SVN). Fig. 3

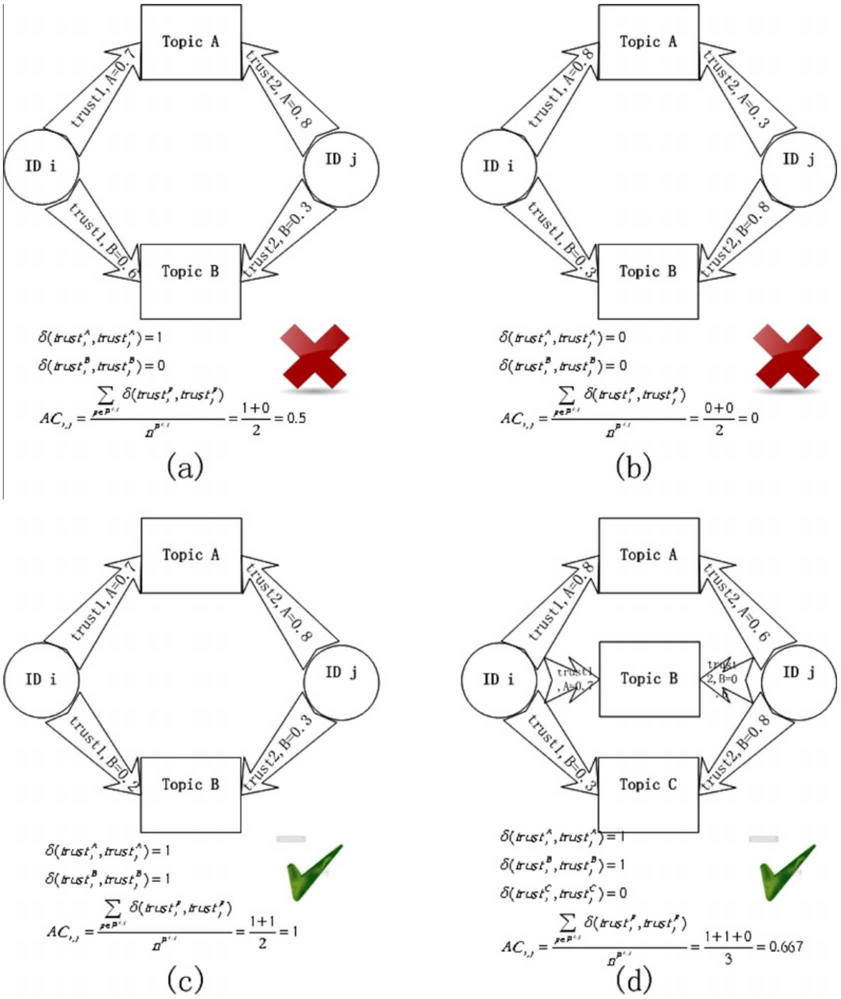

Topic A
trust1, A=0.7
trust2, A=0.8
ID i
ID j
trust1, B=0.6
trust2, B=0.3
Topic B
δ(trust_i^A, trust_j^A) = 1
δ(trust_i^B, trust_j^B) = 0
AC_{i,j} = Σ_{p∈P^{i,j}} δ(trust_i^p, trust_j^p) / n^{P^{i,j}} = (1+0)/2 = 0.5
(a)
Topic A
trust1, A=0.8
trust2, A=0.3
ID i
ID j
trust1, B=0.3
trust2, B=0.8
Topic B
δ(trust_i^A, trust_j^A) = 0
δ(trust_i^B, trust_j^B) = 0
AC_{i,j} = Σ_{p∈P^{i,j}} δ(trust_i^p, trust_j^p) / n^{P^{i,j}} = (0+0)/2 = 0
(b)
Topic A
trust1, A=0.7
trust2, A=0.8
ID i
ID j
trust1, B=0.2
trust2, B=0.3
Topic B
δ(trust_i^A, trust_j^A) = 1
δ(trust_i^B, trust_j^B) = 1
AC_{i,j} = Σ_{p∈P^{i,j}} δ(trust_i^p, trust_j^p) / n^{P^{i,j}} = (1+1)/2 = 1
(c)
Topic A
trust1, A=0.8
trust2, A=0.6
ID i
trust1, A=0.7
Topic B
trust2, B=0
ID j
trust1, B=0.3
trust2, B=0.8
Topic C
δ(trust_i^A, trust_j^A) = 1
δ(trust_i^B, trust_j^B) = 1
δ(trust_i^C, trust_j^C) = 0
AC_{i,j} = Σ_{p∈P^{i,j}} δ(trust_i^p, trust_j^p) / n^{P^{i,j}} = (1+1+0)/3 = 0.667
(d)

Fig. 2. Examples to illustrate the attitudes of a pair of IDs to multiple topics.

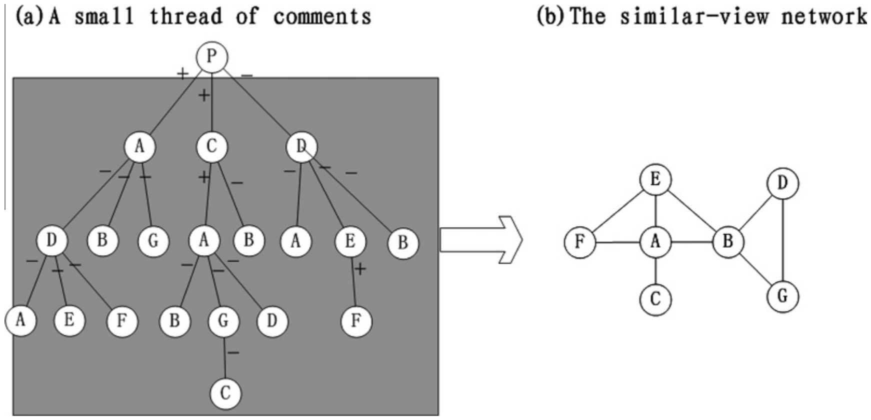


Fig. 3. An example to illustrate the SVN.

gives a simple example to SVN model. As we can see in Fig. 3b, user A and user E are connected, because they reply to user D together, and their attitudes to user D are both negative. The other edges may be deduced by analogy.

4.2. Sock-puppet network

Previous studies indicate authors have their own unique writing style similar to the biological fingerprints, at least for texts produced by authors who are not consciously changing their style of writing across texts [16]. Thus, researchers believe that a set of writings from one author can exhibit greater similarity in terms of these features than ones from different authors. That is to say, any two authors and their works can be distinguished on the basis of their writing styles because of the unique stylistic characteristic for any author. In our similar-view network (SVN), if those connected IDs have similar writing styles, they are more likely from the same person, ally or company; otherwise, they are self-reliant. Therefore, there is a need to add authorship-identification techniques into sock puppet detection task.

4.2.1. Writing-style features

As discussed earlier, the writing-style feature set may significantly affect the performance of authorship identification. Previous studies and analysis have summarized some special characteristics of online comments. Zheng et al. [9] expanded De Vel et al.'s efforts, and integrated four types of features, including lexical, syntactic, content-specific, and structural features.

Lexical features, including character-based lexical features [7,17], vocabulary richness features [18], and word-length frequency features [4,7].
Syntactic features, including function words, punctuation, and part of speech. Those features can capture an author's writing style at the sentence level.
Structural features, those features represent the way an author organizes the layout of a piece of writing.
Content-specific features, including special words or characters closely related to specific topics. The selection of such features is dependent on specific application domains.

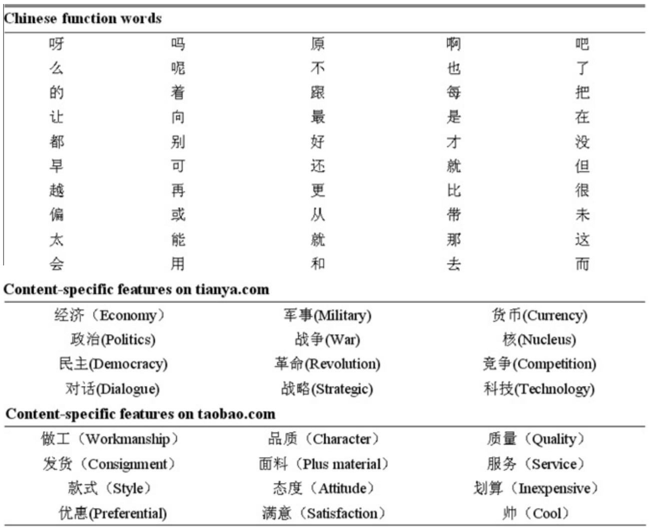
In our work, we investigate the writing-style features on tianya.com and taobao.com to examine the capability of our approach in sock puppet detection. Different from Latin languages such as English and Greek, Chinese is a typical oriental language. For example, some English features (e.g., frequency of the 26 different English letters or average sentence length in terms of word) do not exist in Chinese. What is more, some structural features are difficult to extract accurately in online comments because of their limited length. Here, we remove some features, and select 80 features, including 10 lexical features (Features 54, and 59–67 in [9]), 58 syntactic features (Features 88–95 in [9] and 50 Chinese function words), and 12 content specific features. Function words and content-specific features for the Chinese dataset are listed in Table 2.

4.2.2. Hypothesis testing model

To accomplish the authorship identification task, many analytical approaches have been adopted so far. The two most commonly used techniques are statistical analysis and machine learning approaches. Here, we adopt a statistical model: T-test. Some classification approaches such as C4.5 decision tree [19], back propagation neural networks [20], and SVM [21] can also be used for authorship identification. We choose *t* T-test because of its simplicity and validity. T-test known as the test of statistical significance is a kind of statistical inference method which is used to determine the difference between samples or sample and population.

Definition of T-test[31]: T-test is a two sample locations test of the null hypothesis that the means of two normally distributed

Table 2
Function words and content-specific features.



Chinese function words				
呀	吗	原	啊	吧
么	呢	不	也	了
的	着	跟	每	把
让	向	最	是	在
都	别	好	才	没
早	可	还	就	但
越	再	更	比	很
偏	或	从	带	未
太	能	就	那	这
会	用	和	去	而

Content-specific features on tianya.com		
经济（Economy）	军事(Military)	货币(Currency)
政治(Politics)	战争(War)	核(Nucleus)
民主(Democracy)	革命(Revolution)	竞争(Competition)
对话(Dialogue)	战略(Strategic)	科技(Technology)

Content-specific features on taobao.com		
做工（Workmanship）	品质（Character）	质量（Quality）
发货（Consignment）	面料（Plus material）	服务（Service）
款式（Style）	态度（Attitude）	划算（Inexpensive）
优惠(Preferential)	满意（Satisfaction）	帅（Cool）

populations are equal. All such tests are usually called Student's *t*-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal; the form of the test used when this assumption is dropped is sometimes called Welch's *t*-test. These tests are often referred to as "unpaired" or "independent samples" *t*-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.

Assume there are two samples following normal distribution:

$$X(e) = (e_1, e_2 \cdots e_{n1}) \sim N(\mu_1, \sigma^2) \tag{4}$$

$$Y(e) = (e_1, e_2 \cdots e_{n2}) \sim N(\mu_2, \sigma^2) \tag{5}$$

where e_i is the normalized feature vector, which is calculated by counting the usage frequency of selected function words and content-specific features (Table 2) in a comment.

The process of T-test

```
Input: Two comment samples
Output: Accept null hypothesis or not
  1: Assume the null hypothesis $H_0$:$\mu_1 = \mu_2$ that there is
     really no significant difference between these two independent
     texts, and the alternative hypothesis $H_1$:$\mu_1 \neq \mu_2$;
  2: Given significance level $\alpha$, and sample capacity $n_1$,$n_2$;
  3: Select T-test method, we can get the corresponding statistics $T$.
     $$T = \left| \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|,$$
  where $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e}) \cdot (e_i - \bar{e})'}$ and $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$;
  4: If $T < T_{\alpha/2}(n_1 + n_2 - 2)$, we accept the null hypothesis $H_0$.
     That is to say, there is no significant difference between
     these two samples. Otherwise, we reject $H_0$.
  5: Return: Accept null hypothesis or not
```

4.2.3. Edges pruning

As discussed in Section 3, one's sock puppets may have three interactive characteristics: (1) they often appear in the same theme and reply to same comments; (2) their basic perspectives to a certain topic are often consistent; (3) their writing-style features are similar. The similar-view network built in Section 4.1 obeys the first two characteristics; we need to prune edges whose writing-style features are not similar from SVN. The pruning based on F-test is described as Algorithm 1.

Algorithm 1. Edges pruning based on F-test

```
Input: Original similar-view network
         The writing-style feature set,
         The significance level $\alpha$,
Output:Sock-puppet network
  1: Select the unmarked vertex pair ($v_i$, $v_j$) from SVN, their
     comment sets are extracted from dataset;
  2: Call T-test;
  3: If there is significant difference between these two
     samples, we prune the corresponding edge, otherwise,
     we reserve it;
  4: Mark the vertex pair ($v_i$, $v_j$);
  5: Repeat from step 1 to step 4 until all the vertex pairs
     are marked.
  6: Return: Sock-puppet network
```

The pruned SVN can be called sock-puppet network (SPN). Some structural properties of the obtained network are listed in Table 3, and we also compare it with some existing models. The connectivity of the network (row 3) is the ratio of actual links M to the potential number of links (The number of the edges of a complete graph). As shown in Table 3, SPN is highly sparse compared to the others. In SPN, the "giant component" comprises 65.17% of the users. In SPN, $\langle k \rangle$ is low, meaning that users in this network have a relatively small circle of friends and resemble each other in their interaction behaviors. The clustering coefficient of a node is defined as $C_i = \frac{2E_i}{k(k_i-1)}$, and the clustering coefficient of the whole network is the average of the individual C_i' [25]. We observed that, for SPN, C is much higher than the randomized counterpart which is defined as $C_{rand} = \langle k \rangle / N$. The average shortest path length is small for SPN, suggesting that it is a "small-world" network. The diameter D of this social network is also very small. This has also been seen in other traditional social networks. Another statistic of social networks is the degree correlation, or mixing coefficient, that indicates whether highly connected users are preferentially linked to other highly connected users. Table 3 shows the correlation coefficient r [23,24] (also called the Pearson correlation coefficient) for our four networks. Interestingly, unlike traditional social networks, which exhibit significant assortative mixing, SPN is characterized by disassortative mixing.

4.3. Sock puppet detection based on link analysis

As with other social networks, the sock-puppet network is generally globally sparse yet locally dense. It has vertices in a group structure, where the vertices within the group have a higher edge density, and the vertices between groups have a lower edge density. This kind of structure is called a community, which is an important network property and can reveal many hidden features of a given network. IDs belonging to the same community are likely to have properties in common (e.g., similar writing styles, consistent viewpoints to certain topics, etc.). Those IDs may be the sock puppets from one's self, ally or company. Community identification is a fundamental step not only for discovering what causes entities to form but also for understanding the overall structural and functional properties of a large network.

The methodology for finding these latent communities within networks comes from the physics community and is based on the use of deterministic algorithms. These algorithms focus on optimizing an energy-based cost function that is always defined with fixed parameters over possible community assignments of nodes [26,27]. A notable work proposed by Newman and Girvan [27] introduce modularity as a posterior measure of network structure. Modularity measures interconnectivity and non-interconnectivity; this metric has been influential in the community-detection literature and has found success in many applications [10–15]. Modularity Q is defined as $Q = \sum_i (e_{ii} - a_i^2) = TrE - \|E^2\|$[27], in which E is a $n \times n$ symmetric matrix whose element e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j, and $\|E^2\|$ indicates the sum of the elements of the matrix E^2. The trace of this matrix $TrE = \sum_i e_{ii}$ is the fraction of edges in the network that connect vertices in the same community, while the row (or column) sums $a_i = \sum_j e_{ij}$ give the fraction of edges that connect to vertices in community*i*. If the network is such that the probability to have an edge between two sites is the same regardless of their eventual belonging to the same community, one would have $e_{ij} = a_i a_j$. The modularity measures the degree of correlation between the probability of having an edge joining two sites and the fact that the sites belong to the same community.

To accomplish the sock detection task, we adopt three popular, yet classical community detection algorithms: GN, GIrvan

Table 3
Statistics of the Tianya social networks. Und. Dense, Vicenc Gomez et al. [22]; Und. Sparse, Vicenc Gomez et al. [22]; Semantic, Xia and Bu [15].

	SPN ($\alpha = 0.05$)	SVN	Und. Dense	Und. Sparse	Semantic
N	47,867	261,276	323,745	12,047	162,747
M	132,987	63,891,213	2,987,953	17,680	678,189
Connectivity (%)	0.01161	0.19	0.00057	0.0244	0.00051
Maxclust (%)	65.17	89.19	99.93	85.41	99.26
$\langle k \rangle$	5.56	489	18.46	2.94	8.33
C	0.1134	0.2912	0.0712	0.0287	0.0086
l	3.967	3.12	3.7781	5.29	4.2119
D	10	9	10	17	11
r	−0.0519	0.1005	−0.0899	−0.1285	−0.0760

and Newman [28]; CNM, Clauset et al. [29]; DA, Duch and Arenas [30].

4.4. Summary and analysis of our approach

The overall detection approach is summarized as Algorithm 2. In this section, we explain why we adopt these three main steps.

Algorithm 2. Sock puppet detection based on F-test and link analysis.

```
Input: Social network data set (One theme discussion or
  More),
        The writing-style feature set,
        The significance level $\alpha$,
Output: Sock-puppet communities identified
  1: Construct similar-view network using given social network data set;
  2: Call edges pruned algorithm (Algorithm 1);
  3: Call classical modularity maximization method (GN, CNM, DA).
  4: Return: Sock-puppet communities identified
```

We analyze the time complexity of our approach by considering the two major computational steps: the edges pruned algorithm (step 2) and classical modularity maximization method (step 3). In the step 2, the main influence on the time complexity comes from feature extractor for every comment, the time complexity of this process is $O(A \cdot C)$, where A is feature number and C is the total number of comments respectively. Then edges are analyzed for pruning, its time complexity is $O(M^V)$, where M^V is the number of edges in SVN. As to the modularity maximization based method, the time complexity of three classical community-detection algorithms is listed in Table 4. The parameter d in GNM [29] is the depth of the dendrogram describing the community structure. Therefore, the time complexity of our approach can reach $O(A \cdot C + M^V + N^P \log^2 N^P)$, where N^P is the number of nodes in SPN. As database technology has advanced, feature extractor for every comment can be real-time synchronous. It will greatly simplify our work: the overall time complexity is reduced to $O(M^V + N^P \log^{2\text{-}} N^P)$. On the sparse network, our approach runs in essentially linear time.

Table 4
Comparison of the time complexity for the network division found by classical methods. GN, Glrvan and Newman [28]; CNM, Clauset et al. [29]; DA, Duch and Arenas [30].

	GN	GNM	DA
Arbitrary graph	$O(M^2N)$	$O(Md\log N)$	$O(N^2\log N)$
Sparse graph	$O(N^3)$	$O(N\log^2 N)$	$O(N^2\log N)$

Compared to traditional methods, our approach has three advantages: (1) it conforms to the practical meanings of sock puppet community; (2) it can be applied in online situation; (3) it increases the efficiency of link analysis.

As shown in Fig. 4, we assume that the real hosts of each ID have been known and labeled in Fig. 4a. That is, IDs A, E and F are from person P1; IDs B, D and G are from person P2; ID C is from person P3. We notice in Fig. 4a that, IDs from the same person often reply to the same topic or comment, and their standpoints are strikingly similar. For instance, IDs B, D and G together comment to ID A twice, and their attitudes are negative. To grasp this property, we build the similar-view network as shown in Fig. 4b. Then, we add authorship-identification techniques into our approach, because one's writing style is steady-going. Fig. 4c shows a sock-puppet network in which connected IDs have similar writing style. Finally, we apply classical modularity maximization method to the sock puppet detection task, and get communities we expect as shown in Fig. 4d. IDs belonging to the same community have properties in common: similar writing-style, consistent viewpoint to a certain topic. It conforms to the practical meanings of sock puppet community.

Another advantage of our approach is that it can be applied in online situation. Traditional authorship identification techniques [3,9,34] mainly focus on the identification of a limited number of articles or texts. Compared to realistic environment, the online situation is further complicated by the sheer amount of cyber users and activities. The articles or texts in the online situation are related according to users' activities. So, massive amounts of the real data collected from online societies are network structured. Our approach tactfully uses the interactive relationship between users as the "navigation information". With advanced database technology, it can prove the feasibility of sock puppet detection task in online situation.

Finally, our approach increases the efficiency of link analysis. As the number of connections between the post's author and its direct commentators may be very large, the obtained similar-view network is a dense graph. On the one hand, the community structure of a dense graph may not be significant; the time complexity of classical modularity maximization methods to a dense graph is very high on the other. Performing edges pruning can largely decrease those unrelated links (see Table 3), it can reduce the complexity of link analysis. Moreover, the community structure of sock-puppet network is more significant.

5. Experiment and analysis

We conducted an experimental study to evaluate our proposed approach; two real datasets used in our experiments are described in the following:

Tianya dataset: Tianya forum (http://focus.tianya.cn) is a popular bulletin-board service in China. It includes more than 300 boards, and the total number of registered user identifications

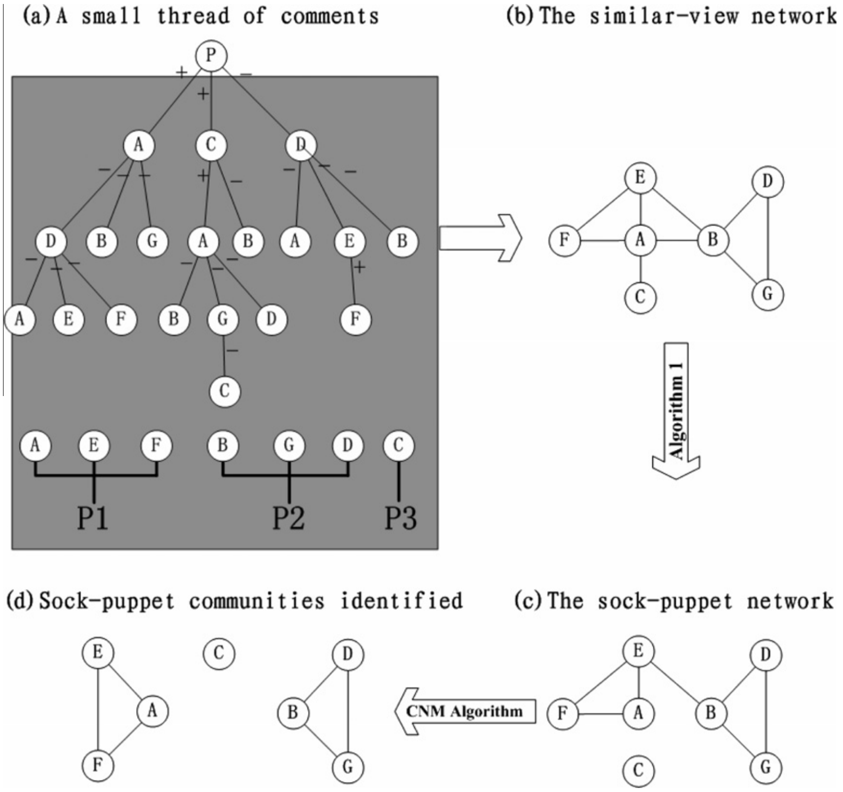
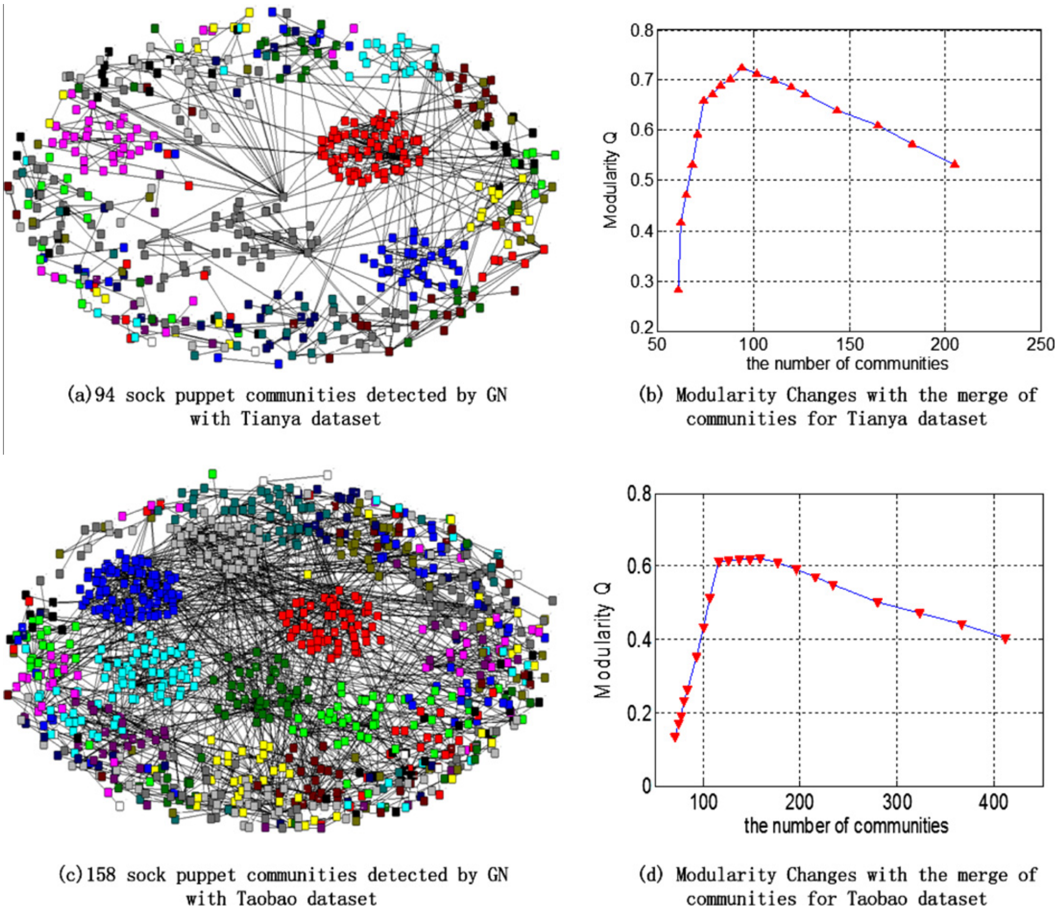

(a) A small thread of comments
(b) The similar-view network
P
+
+
−
A
C
D
D
B
G
A
B
A
E
B
A
E
F
B
G
D
F
C
A
E
F
B
G
D
C
P1
P2
P3
E
D
F
A
B
C
G
Algorithm 1
(d) Sock-puppet communities identified
(c) The sock-puppet network
CNM Algorithm

Fig. 4. An example to illustration the advantage of our approach.



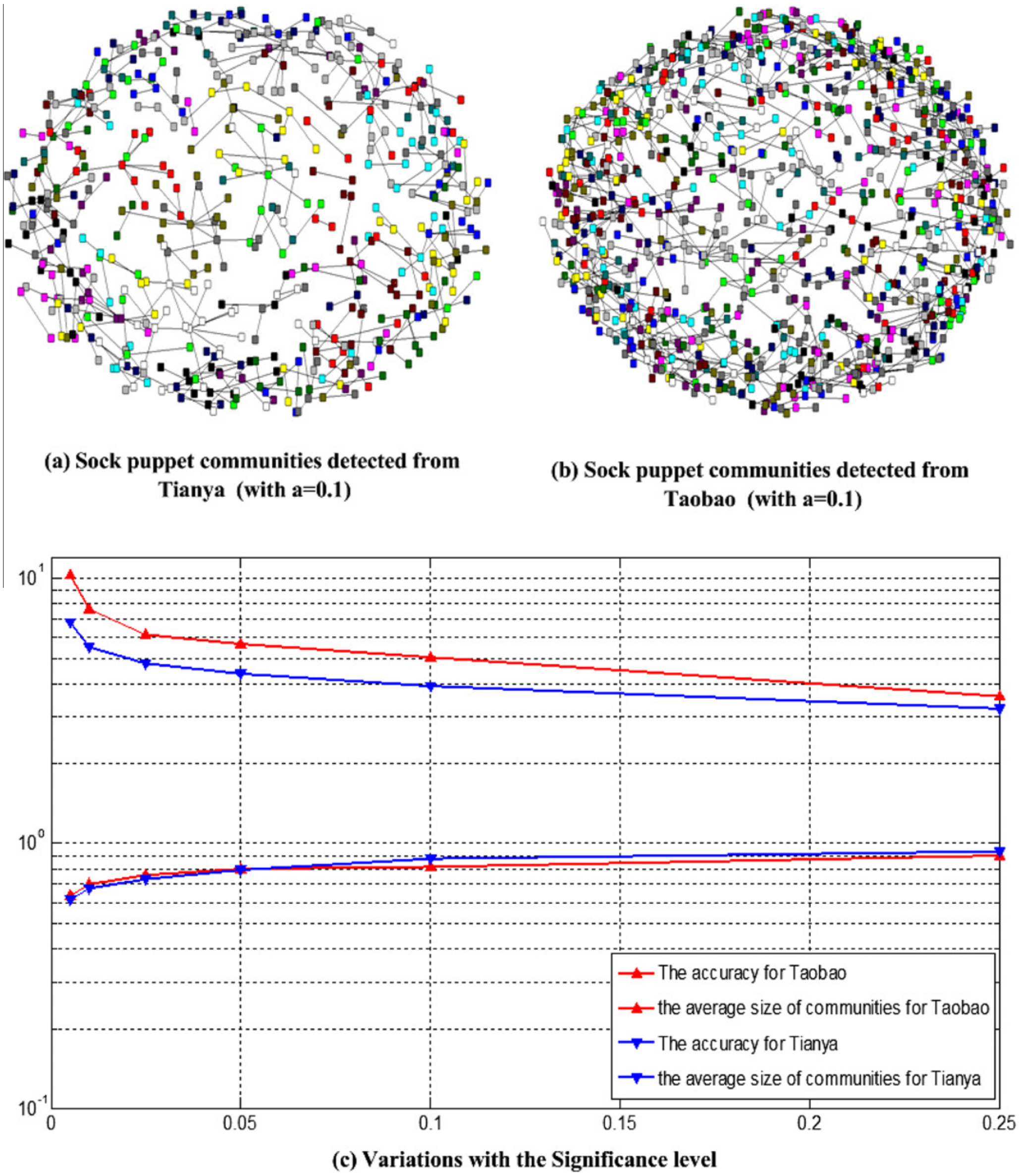
(a) 94 sock puppet communities detected by GN with Tianya dataset

(b) Modularity Changes with the merge of communities for Tianya dataset

(c) 158 sock puppet communities detected by GN with Taobao dataset

(d) Modularity Changes with the merge of communities for Taobao dataset

Fig. 5. Sock puppet communities detected from two datasets (with $\alpha = 0.05$), different colors are used to represent different communities.



(a) Sock puppet communities detected from Tianya (with a=0.1)

(b) Sock puppet communities detected from Taobao (with a=0.1)

(c) Variations with the Significance level

Fig. 6. The accuracy and average size of communities change with the significance level α.

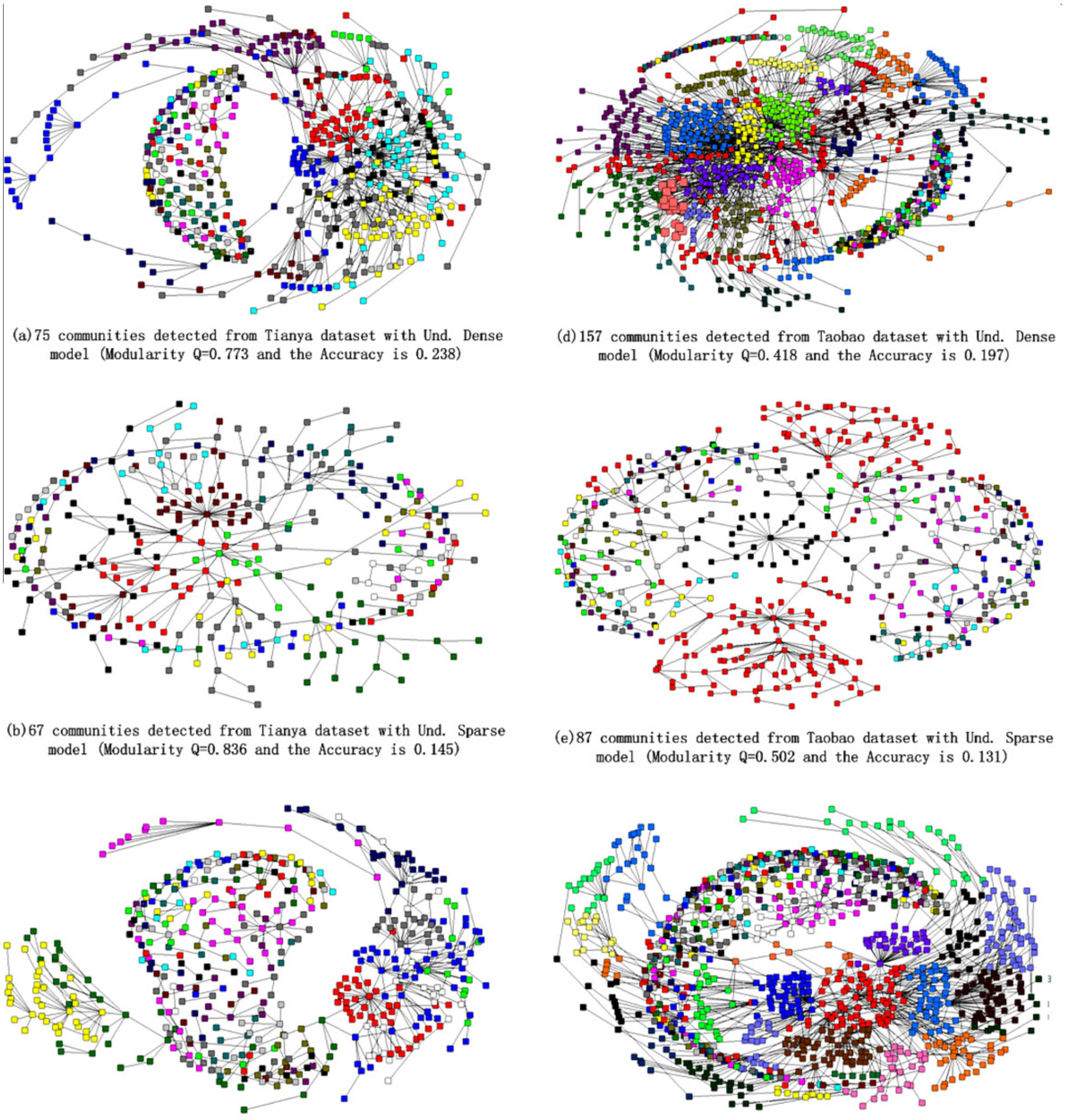
(IDs) is more than 32 million. Since its introduction in 1999, it has become the leading social-networking site in China due to its openness and freedom. We selected the worldview board and collected data between July, 2003 and December, 2011. Here, we puck up a subset of Tianya dataset including 539 users, 11 threads and 4951 replies.

Taobao dataset: Taobao (http://www.taobao.com) is an ebay like website founded in middle 2003. It is the largest online auction and shopping website in which people and businesses buy and sell a broad variety of goods and services worldwide in Asia. We download commercial evaluation information from customers including 127 shop home pages and 9854 comments.

5.1. Overall performance

To analyze the above two datasets, we first constructed similar-view networks using the method introduced in Section 4.1. Here, the subset of Tianya includes 539 users and 4951 replies (average reply number is 9.19); while, the Taobao dataset includes 980 users and 9845 comments (average reply number is 10.05). As these two average values are around 10, we determined the default value of this threshold ϕ as 10. Given the writing-style feature set (see Table 3) and the significance level ($\alpha = 0.05$), we called Algorithm 1 to prune those weak-related links. Finally, we detected sock puppet communities for each SPN with classical community detection algorithm. The results are shown in Fig. 5.

With the Tianya dataset, the board we selected is worldview which is about international news and current affairs. With our sock puppet detection approach, we obtained 94 communities with the average size of 4.381, as shown in Fig. 5a; here, the modularity Q is 0.7245, which is a very large figure. This means that community character is obvious in sock-puppet network. We noticed that there are many isolated nodes which are grouped mainly in pairs or, at most, in small clusters of four, those IDs may be sock puppets from one' self; there are still some larger communities, hosts behind those IDs may from the same ally or company. For the Taobao dataset, Fig. 5c shows the sock puppet communities detected from online message boards of 127 shops. We detected 158 communities with the modularity Q of 0.621. Here, the average size of communities is 5.644, and the minimum size is 2, the



(a) 75 communities detected from Tianya dataset with Und. Dense model (Modularity Q=0.773 and the Accuracy is 0.238)

(d) 157 communities detected from Taobao dataset with Und. Dense model (Modularity Q=0.418 and the Accuracy is 0.197)

(b) 67 communities detected from Tianya dataset with Und. Sparse model (Modularity Q=0.836 and the Accuracy is 0.145)

(e) 87 communities detected from Taobao dataset with Und. Sparse model (Modularity Q=0.502 and the Accuracy is 0.131)

(c) 95 communities detected from Tianya dataset with Semantic model (Modularity Q=0.878 and the Accuracy is 0.445)

(f) 178 communities detected from Taobao dataset with Semantic model (Modularity Q=0.531 and the Accuracy is 0.397)

Fig. 7. Communities identified based on traditional models (Und. Dense, Und. Sparse, and Semantic) within two datasets, different colors are used to represent different communities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

maximum size is 52. That means the users in Taobao are more likely grouped in unified organizations. The main cause leading to this is that the sellers in Taobao are driven by self-interests. For instance, to believe that a product is a good buy, those sellers may use different fake online identities pretending to be different persons to praise or create the illusion of support for the product.

5.2. Evaluation criterion

Since there is no real answers, for each reported sock puppet community, we examined all their posts and replies and based on statistical analysis to determine if they are sock puppets or not. Some other traces can be used for sock puppet detection in Tianya and Taobao:

(a) Login IP: If the two IDs have similar login IP addresses (i.e., the first three bytes of their IP addresses are the same), we can doubt that one of them is the sock puppet of the other's;

(b) ID: If the two IDs have similar spellings or pronunciations (i.e., one ID is called "Vigi", the other is called "Vigian"), they are both suspicious;

(c) Avatar: We can also check whether these two IDs have similar avatars;

(d) Registration time: If their registration time are very close (i.e., one's registration time is "2010-2-12", the other's is "2010-2-16"), we can also doubt their

(e) truthfulness;

(f) Last login time: One's sock puppet often appears in company with its host ID; therefore, we can check whether their last login times are on the same day.

Therefore, in this paper, if the detected ID pair with our method obeys above three or more conditions, we believe that it is the correctly detection. To verify the above detection traces, a questionnaire investigation is sent to all the detected ID pairs, and feedback information has confirmed our judgments. For example, we selected a five node sock puppet community from Fig. 5a. We looked up the contents that were posted by the community from the database.[1,2] On December 17th 2011, one user called "Selinazp" posted a topic entitled "Strong Chinese Navy". After several minutes, a user named "furisa" pushed the topic and said Selinazp's thinking is rust and freeze. Then, the other two users called "Wangfir" and "Bowisn" rebutted furisa's comment a short time later. After about 15 min, "Tomwar" and "Joinfish" joined in and supported the author. Almost at the same moment, Joinfish posted another topic entitled "The power projection of Chinese Navy". By coincidence, Wangfir, Bowisn and Selinazp replied to this topic and expressed their similar viewpoints. The activities of those five users were strikingly parallel, moreover, their registration time on Tianya are in the same day. Analyzing their writing styles, we found that their usages of function words are similar more or less. So it is quite certain that Wangfir, Bowisn, Tomwar and Joinfish were in collusion with Selinazp to drive up the popularity of his thesis. The similar examples can be found in other detected community in Fig. 5a.

The accuracy of our approach was defined as C/M^P, where M^P is the number of edges in SPN, and C is the total number of correctly detected pairs in SPN. In fact, the accuracy of our approach changes with the topology of SPN, and the later changes with the significance level α. Therefore, we would expect that the accuracy of our approach also changes with the level α. Then, we discuss the relationship between the accuracy and the significance level α in detail. With the same dataset, we repeated the experiments; the significance level α was varied from 0.005 to 0.25. As shown in Fig. 6, as α grew, the accuracy was declined, whereas, the average size of communities grew fast. This means that sock puppets from the same alliance or company may not be detected with low significance level. So, we need to set up an appropriate significance level in our approach.

5.3. Comparison

In this subsection, we compared our model with some existing models [15,22]. We used the metric defined in Section 5.2 and the largest modularity value Q to evaluate the experimental results. With the Tianya dataset, there are 75 communities in Und. Dense Network as shown in Fig. 7a. The modularity of its community structure is 0.773 which is very high, however, the accuracy is only 0.238. In the Und. Spare Network as shown in Fig. 7b, the modularity is higher, which reaches 0.836, but the accuracy of detected sock puppets is very low. Then we tested our model, the result is displayed in Fig. 7c. We found that both the modularity and the sock puppet accuracy are higher than the former models. For the Taobao dataset, we took the same steps. We firstly examined the modularity and the accuracy in Und. Dense Network; as shown in Fig. 7d, the former is 0.418 and the later is 0.197, both of which are very low. Then, we built the Und. Sparse Network, Fig. 7e shows the similar characteristic as Fig. 7d. Finally, a sock-puppet network was constructed, the modularity reaches 0.531 which is the highest for the given dataset and the accuracy is 0.397 which is also higher than the former two (Fig. 7f). With our model, the constructed network has better community structure, and the accuracy of detected sock puppets is higher. Therefore we can conclude that our approach has a better performance than existing models.

6. Conclusion

Online virtual networks provide an excellent platform for users to communicate and share knowledge. On the other hand, it also facilitates cyber deceptions, such as users' ID theft, article counterfeit, and swindle. One common trick to cheat people to believe fake products or a high-return low-risk investment scheme in the online virtual networks is to make use of sock puppets.

In this paper, we propose a sock puppet detection algorithm which combines authorship-identification techniques and link analysis. Firstly, we propose an interesting social network model (SVN) in which links between two IDs are built if they have similar attitude to most topics both of them participate in. Then, the edges are pruned according to a hypothesis testing: (1) Given two IDs who are connected in the network, particular comment sets are extracted from dataset respectively to check their writing features; (2) the null hypothesis is that these two sets are from the same person; (3) the test value T is calculated; (4) in a given test level α, if the null hypothesis is true, we reserve their edge, otherwise, we remove it. The pruned SVN is called sock-puppet network (SPN). Finally, the link-based community detection for SPN is performed. The algorithm is efficient that it can be applied to real dataset. From our experiments, we find the results are promising.

Acknowledgments

This work was supported by Funding of Jiangsu Innovation Program for Graduate Education (Project No: CXZZ12_0162), the Fundamental Research Funds for the Central Universities and Shanghai Intelligent Information Processing Laboratory (Subject No: IIPL-2011-007).

References

[1] V. Akre, A.H. Rizvi, M. Arif, Online social networks – an interface requirements analysis, in: 2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 550–556.

[2] H. Memic, A, Joldic, A more comprehensive activity analysis of standard online social networking functionalities, in: 2nd International Conference on Software Technology and Engineering (ICSTE), 2010, pp. V2-108–V2-111.

[3] Abbasi Ahmed, Chen Hsinchun, Applying authorship analysis to Arabic Web Content, Lecture Notes in Computer Science 3495/2005 (2005) 75–93.

[4] T.C. Mendenhall, The characteristic curves of composition, Science 11 (11) (1887) 237–249.

[5] F. Mosteller, D.L. Wallace, Applied Bayesian and Classical Inference: The Case of The Federalist Papers, Springer Series in Statistics, 1964.

[6] F. Mosteller, D.L. Wallace, Inference and Disputed Authorship: The Federalist, Addison-Wesley, Reading,Mass, 1964.

[7] O. De Vel, Mining E-mail authorship. Workshop on Text Mining, in: ACM International Conference on Knowledge Discovery and Data Mining (KDD 2000).

[8] O. De Vel, A. Anderson, M. Corney, G. Mohay, Mining email content for author identification forensics, 30(4) (2001) 55–64.

[9] Zheng Rong, Li Jiexun, Chen Hsinchun, Huang Zan, A framework for authorship identification of online messages: writing-style features and classification techniques, Journal of the American Society for Information Science and Technology 57 (3) (2006) 378–393.

[10] J. Zeng, S. Zhang, C. Wu, A framework for WWW user activity analysis based on user interest, Knowledge-Based Systems 21 (8) (2008) 905–910.

[11] N. Du, B. Wu, X. Pei, B. Wang, L. Xu, Community detection in large-scale social networks, in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, ACM, 2007, pp. 16–25.

[12] A. McCallum, A. Corrada-Emmanuel, X. Wang. Topic and role discovery in social networks, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005, pp. 786–791.

[13] Y. Tian, R. Hankins, J. Patel, Efficient aggregation for graph summarization, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, pp. 567–580.

[14] Zhongying Zhao, Shengzhong Feng, Qiang Wang, Joshua Zhexue Huang, Graham J. Williams, Jianping Fan, Topic oriented community detection through social objects and link analysis in social networks, Knowledge-Based Systems (26) (2012) 164–173.

[15] Z. Xia, Z. Bu, Community detection based on a semantic network, Knowledge-Based Systems (26) (2012) 30–39.

[16] D.I. Holmes, Authorship attribution, Literary and Linguistic computing 13 (3) (1998) 111–117.

[17] R.S. Forsyth, D.I. Holmes, Feature finding for text classification, Literary and Linguistic Computing 11 (4) (1996) 163–174.

[18] F.J. Tweedie, R.H. Baayen, How variable may a constant be? Measures of lexical richness in perspective, Computers and the Humanities 32 (1998) 323–352.

[1] http://www.tianya.cn/publicforum/content/worldlook/1/419049.shtml

[2] http://www.tianya.cn/publicforum/content/worldlook/1/419047.shtml

[19] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[20] R.P. Lippmann, An introduction to computing with neural networks, IEEE Acoustics Speech and Signal Processing Magazine 4 (2) (1987) 4–22.
[21] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.
[22] Vicenc Gomez, Andreas Kaltenbrunner, Vicente Lopez, Satistical analysis of the social network and discussion threads in Slashdot, in: WWW 2008, April 21–25, 2008, Beijing, China.
[23] M.E.J. Newman, Assortative mixing in networks, in: PhysRevLett 89, 208701.
[24] M.E.J. Newman, Mixing patterns in networks, Physical Review E 67 (2003) 026126.
[25] E.M. Trevino, Blogger motivations: power, pull, and positive feedback, in: Internet Research 6.0, 2005.
[26] J. Reichardt, S. Bornholdt, Statistical mechanics of community detection, Physical Review E 74 (2006).
[27] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2004).
[28] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences USA 99 (2002) 7821–7826.
[29] A. Clauset, M.E.J. Newman, C. Moore, Finding community structure in very large networks, Physical Review E 70 (2004) 066111.
[30] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Physical Review E 72 (2005) 027104.
[31] http://en.wikipedia.org/wiki/T-test.
[32] Richard Morin (February 1, 2003), Scholar Invents Fan To Answer His Critics, Washington Post <http://www.washingtonpost.com/ac2/wp-dyn/A8884-2003Jan31> (retrieved 5.6.09).
[33] Hari A. Johann, Personal Apology, The Independent (website), 14 September 2011; Richard Seymour, ‘‘The Johann Hari Debacle, The Guardian, September 16, 2011.
[34] Ahmed Abbasi, Hsinchun Chen, Visualizing authorship for identification, LNCS 3975 (2006) 60–71.
[35] http://en.wikipedia.org/wiki/Sockpuppet(Internet).