Measuring Domain Influence in Heterogeneous Networks

Quan Liu liuquan@pku.edu.cn
Chenguang Wang wangchenguang@pku.edu.cn
Ming Zhang mzhang@net.pku.edu.cn
School of EECS, Peking University, No.5 Yiheyuan Road, Haidian District, Beijing 100871, China

ABSTRACT
Influence can greatly benefit fields like viral marketing, information propagation and recommender systems, while the pervasiveness of heterogeneous networks, such as Twitter, provides richer information for influence research. However, current influence research focuses on analyzing general influence, which assumes various users have similar influence over the network, without fully exploiting the rich information in underlying heterogeneous networks; while in real world, users always belong to specific domains based on these heterogeneous information (e.g. relations and interested topics). In order to enhance influence analysis by providing subtle domain-level influence view, in this paper, we present a systematic approach modeling domain influence in heterogeneous networks. We first utilize spectral clustering to generate partitioned domains based on a three-dimension heterogeneous network, which includes directional relations (i.e. followers and followees) and topics. Then we measure domain influence respectively by widely used metrics: No. of Followers, No. of Retweets, and PageRank. The experiment is conducted on the real world dataset Sina Weibo (a famous Chinese microblog). The results indicate that there is a stronger correlation between different measures in domain influence than general influence, especially when the domain is highly specialized, with the best Spearman' correlation coefficient gaining 0.42 (0.9972 – 0.5782) and Kendall' tau gaining 0.39 (0.9720 – 0.5868).

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database application– Data mining; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – Information filtering

General Terms
Experimentation, Human Factors, Measurement

Keywords
Microblogs, Domain Influence, Heterogeneous Network

1. INTRODUCTION
Influence plays a key role in viral marketing, recommender systems and many other fields. Ever since the booming of heterogeneous online social networks, such as Twitter and Weibo, richer information is available for influence research. Yet current research focuses on analyzing users' influence under the assumption that a user's influence persisting the same no matter whom the influence will affect. Besides, the rich information provided by underlying network is not fully explored; however, in real world, users participate online with numerous activities, such as following, tweeting, etc. As a result, users have heterogeneous properties. Following forms different relations, tweeting leads to various topics. Then, different users can form different domains based on their heterogeneous properties. Thus, domain-based influence analysis is important. Moreover, it provides us a subtle view of influence, and further benefits the related applications.

For example, in Weibo, if user A follows B, B is A's followee. A and B may belong to the same domain; if A is followed by B or A, B has similar tweets, they'll also probably belong to the same domain. Motivated by this, in this paper, we present a systematic approach modeling domain influence in heterogeneous networks. First, we use spectral clustering to generate domains based on a three-dimension heterogeneous network (followers, friends, topics) by applying utility integration [8]. Then, in each domain, we measure domain influence separately by well-known metrics: No. of Followers, No. of Retweets and PageRank. [4]

Related work can be divided into two parts. The first is on modeling influence. Kempe et al. [1] presented two stochastic influence diffusion models: Independent cascade (IC) model and Linear threshold (LT) model. From then on, research has been focusing on validating the existence of influence, and quantifies influence with different measures. Anagnostopoulos et al. [2] proposed a shuffle test to prove the existence of social influence. Timothy et al. [3] presented a randomization technique to measure correlation gain based on influence and homophily. Kwak et al. [4] compared three different measures of influence: No. of Followers, PageRank, and No. of Retweets. Meeyoung et al. [5] also presented three measures, replacing PageRank with No. of Mentions and made several interesting observations.

However, all these work has been focusing on modeling user influence without considering the domain knowledge of each audience of the influence. Yet in real world, it is impossible for a person to excel in every domain. Life experience also reveals that people outside our industry or interest tend to have less impact on us, though they may be quite influential in their own domain. Recently, several efforts have been made to research into topical influence. Jie et al. [6] utilized node-specific topic distribution to analyze the topic-level influence and scale it to large social graphical networks. Lu et al. [7] aimed at the same goal using a probabilistic model and extend it to indirect influence. However, both of them and other relevant topical influence research only employed link information for generating node-specific topics, without further using them as well as content topics to form domains, and measure influence in each domain.

The second is community detection research. A lot of work has been proposed to optimize clustering algorithm in social networks and address the challenge of heterogeneous networks. Our work employs the utility integration method proposed in Lei et al. [8] to combine different dimensions of the heterogeneous network and further use spectral clustering [10] to generate domains. This integration of different utility matrix allows us to gain insights from both structure and content views, which is also a major difference from previous work that uses unilateral information.

Experiment results on dataset Sina Weibo, a very popular microblog website in China, shows the improvement of Spearman's (+0.42) and Kendall's tau (+0.39) rank coefficient, which indicates the growth of correlation between No. of Followers and No. of Retweets as well as PageRank in domain influence than in global influence (without considering domain knowledge). Such improvement compared with previous work [4,

5] further demonstrates the significance of effectiveness when introducing domains into influence research.

This paper is organized as follows. We present our approach to model domain influence in Section 2. Section 3 introduces our experiments on Weibo dataset. At last we conclude in Section 4.

2. Approach

In this section, we present our approach modeling domain influence in heterogeneous networks. The model is divided into two stages: domain discovery and domain influence analysis.

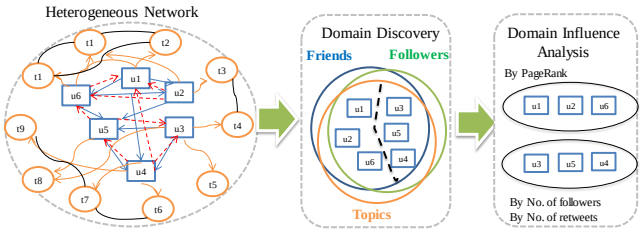

Figure 1. Domain influence modeling

Take Sina Weibo as an example, as shown in Fig. 1 (ui represents user id, ti denotes tweet id), a user has three heterogeneous properties: 1) users follow each other as well as 2) tweet or retweet others' tweets, which are shown as the directed edges with different colors; 3) since tweets share similar content or topics, we use undirected edges to represent their relation. In domain discovery stage, we employ spectral clustering to generate separate domains based on utility integration of each user's heterogeneous properties; In domain influence analysis stage, we rank user influence in a specific domain obtained in stage 1 with three measures: No. of Followers, No. of Retweets and PageRank.

2.1 Heterogeneous Network

On social networks like Twitter or Weibo, users can interact with each other through various media. For example, a user can follow somebody, or retweet his tweets. In order to facilitate later discussion, we propose a few essential definitions and notations.

Define a network G = (V, E;Ω). V is a set of nodes, which are classified into T types,$\Omega = \{X_t\}_{t=1}^{T}$, where X_t is a set of nodes and its edges with the t-th type. The edge set E, a subset of V × V, denotes the connections between nodes. For $\forall\, e_{uv} = (u, v) \in E$, if there exists an edge between u and v, e_{uv} = 1; otherwise e_{uv} = 0. The edges can be either directed or undirected.

In this paper, we gather both structure and content information from Weibo. Since the structure is composed of unidirectional links, for each user u, people he follows and people who follow him are two different dimensions. In other words, we can infer whether two users belong to the same domain based on three pieces of information: (1) whether they have similar followers; (2) whether they follow similar people; (3) whether they are talking about similar topics in their tweets.

Therefore, we propose a three-dimensional heterogeneous network, $\Omega = \{X_t\}_{t=1}^{3} = \{X_1, X_2, X_3\}$, where X_1 is the Friend Network, X_2 the Follower Network, X_3 the Topic Network.

In Friend Network X1 = (V1,E1), e_{uv} = 1 if user u follows user v, 0 otherwise; In Follower Network X2 = (V2,E2), e_{uv} = 1 if u is followed by v, 0 otherwise; while in Topic Network, things get a little more complicated. After word segmentation, all the tweets are input into the Twitter-LDA [9] in order to discover common topics through unsupervised learning. LDA is a famous graphical model for topic discovery while Twitter-LDA further extends it to be more adaptable to the short texts on social networks. Based on the probability distribution output of each vocabulary on each learned topics, we iterate every segmented word of every user and generate a 20-dimension topic vector for each user, forming the Topic Network X3.

2.2 Domain Discovery

To measure domain influence, we need to acquire these different domains out of the heterogeneous network X1, X2, X3 first. Spectral clustering has been shown to be more effective in finding clusters than some traditional clustering algorithms in online social networks, and according to Lei's study [8], it can also fit into the heterogeneous condition because of the equivalency between utility matrix and Laplacian matrix.

Let W denote the weighted adjacency matrix of the similarity graph of each heterogeneous network X1, X2, X3, D the corresponding degree matrix, I the cell matrix. The graph Laplacian matrix L is defined as:

$$L = \begin{cases} D - W & , Ratio\ cut \\ I - D^{-1/2} W D^{1/2}, & Normalized\ cut \end{cases} \quad (1)$$

In this paper we use the normalized cut and compute the first k eigenvectors of L, which decreases n-dimension to k. Three graph Laplacian matrix L1, L2, L3 are generated respectively for Friend Network X1, Follower Network X2 and Topic Network X3.

We employ the utility integration method proposed by Lei et al.[8] to combine the three-dimension network. An average utility matrix can be obtained as follows:

$$\bar{M} = \frac{1}{d} \sum_{i=1}^{d} M^{(i)} \quad (2)$$

where for spectral clustering, the utility matrix M equals the graph Laplacian matrix, so we can derive the average Laplacian:

$$\bar{L} = \frac{1}{3} (L_1 + L_2 + L_3) \quad (3)$$

which makes use of all three networks X1, X2, X3. At last we perform spectral clustering and acquire 5, 10, 15 and 20 communities respectively.

2.3 Domain Influence Analysis

Previous comparison of general influence measures in [4,5] has shown that No. of Followers not related to No. of Retweets, while PageRank is similar to No. of Followers to some extent. With generated domains, we are now able to compare these measures in each domain to quantify domain influence.

We can directly obtain average No. of Retweets from crawled data. As for No. of Followers, we first filter spams and extremely inactive users, then we update it by counting the exact number of followers in the domain that the user belongs to, rather than the general number which was used in previous work.

PageRank is an algorithm first introduced by Google to rank websites in their search engine results. Here we map each user (node) to a website, connections between users to links among those sites. And iteratively distribute a user's influence score to his friends. Experiments show that the convergence is quickly reached no matter what initial scores we set.

With previous computation, we respectively rank users in each domain by the three measurements. Both Spearman's correlation and Kendall's tau are calculated on top of it. In order to compare with general influence over the whole network, we also rank in the universal set. And to avoid the bias of low-influence users who have 0 retweets, we further compute the coefficients on top

20% users. These two coefficients will serve as the baseline. And detailed results will be discussed in the Experiment section.

3. EXPERIMENT

3.1 Dataset and Metrics

We conduct our experiments on a real-world microblog dataset, which is crawled from Sina Weibo (http://weibo.com), a microblog system like Twitter. We select around 100 middle-class seeds (who have tens of thousands of followers) from different industries by Sina human labels, and crawl their 1-level connections. After removing duplicates, we get 1,805,504 users and then crawl their recent 100 tweets (basically between April to May 2013). To avoid noises caused by spam users and extremely inactive users, we perform 4 layers of filtering and reach a steady community of 1,080,204 users and more than 100 million tweets:

(1) Users whose profiles are not available at the moment;

(2) Users who have few followers or friends;

(3) Users who have few tweets;

(4) Users who have few connections in the crawled network.

Another thing to be noted here is the difference between all sorts of online social networks. As for a co-author network like DBLP, their scope is much more focused, where authors are directly connecting those in their field of study. However, for twitter or weibo, they allow weak connections, i.e. unidirectional links like following or being followed; and users' daily activities are more general, which may involve many different aspects. Therefore combining both content and structure information to partition domains is more significant in such scenario. We further repeated our experiment on a relatively small network crawled from Twitter (around 100,000 users) and received similar growth as Sina Weibo. Detailed discussion of Twitter results is skipped due to repetition and limited space.

For correlation analysis, we use Spearman's correlation coefficient and Kendall's tau. Spearman's is defined as follows:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N^3 - N} \quad (6)$$

where xi, yi denote the ranks by two different measures; N is the total number of users. And the Kendall's tau is calculated as:

$$\tau = \frac{2P}{\frac{1}{2}N(N-1)} - 1 \quad (7)$$

where P is the sum over all users, of the number of users ranked after the given user by both rankings.

3.2 Results of Domain Discovery

Using the average Laplacian matrix $\bar{L}$, we perform spectral clustering respectively into 5, 10, 15, 20 domains and measure their user distribution, which is shown in Fig. 2.

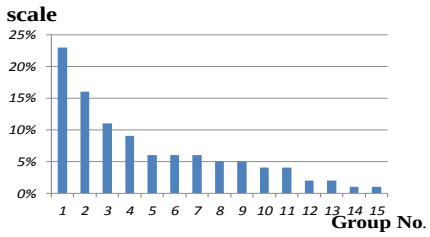

Figure 2. User distributions on 15 clusters

From the above figure we can see that about 3 or 4 communities dominate others, and altogether account for around 50% of the whole network, while the others almost evenly divide another 50%. Results from 5-clusters, 10-clusters and 20-clusters also resemble the figure above, with a few dominating while the rest evenly divided. The granularity, on the other hand, tends to be smaller as the number of clusters grows.

3.3 Measuring Domain Influence

Let $\boldsymbol{\rho_{F-R}}$ represent the Spearman's coefficient between No. of Followers and Retweets; $\boldsymbol{\rho_{F-P}}$ between No. of Followers and PageRank; $\boldsymbol{\tau_{F-R}}$ the Kendall's tau coefficient between No. of Followers and Retweets; $\boldsymbol{\tau_{F-P}}$ between No. of Followers and PageRank. For each community in the 15-clusters' result, we rank users respectively by the three measures and calculate coefficients listed above. Same calculation is also performed on the general influence of 108,204 users. And to avoid the bias of low-influence users who have 0 retweets, we further compute the coefficients on top 20%. Table 1 shows results of some representative domains.

Table 1. Spearman's & Kendall's for domain influence

	All	*Top 20%*	*Domain 1*	*Domain 2*	**Domain3**	*Domain4*	**Domain5**	*Domain6*
ρ_{F-R}	*0.5265*	*0.408*	*0.5495*	*0.6455*	**0.6600(+0.13)**	*0.6375*	**0.7581(+0.23)**	*0.5541*
τ_{F-R}	*0.5761*	*0.3052*	*0.5739*	*0.6231*	**0.6294(+0.05)**	*0.6162*	**0.7018(+0.13)**	*0.5778*
ρ_{F-P}	*0.5782*	*-0.382*	*0.9641*	*0.9718*	**0.9699(+0.39)**	*0.9684*	**0.9972(+0.42)**	*0.967*
τ_{F-P}	*0.5868*	*-0.3163*	*0.8937*	*0.8983*	**0.8946(+0.31)**	*0.8899*	**0.9720(+0.39)**	*0.8918*

In Table. 1, Domain 1 is on economy and company; Domain 2 is a charity community; Domain 3 turns out to be the IT industry; Domain 4 is about literature; Domain 5 is a non-government organization called Lions' Club; Domain 6 is on art.

For Follower-Retweet's correlation, we observe an obvious growth from Baseline I (Top 20%) and Baseline II (Universe) for both Spearman's and Kendall's coefficients. Domain 5 reaches the highest Spearman's value of over 0.7, increasing from the universal baseline by 0.23, which indicates a very strong correlation between No. of Followers and No. of retweets in this domain. In the next section we'll show that this domain (Lion's Club) turns out to be a very dense sub-graph, with users more strongly linked to each other.

As for Follower-PageRank's Correlation, all domains' results outperform the two baseline in a large deal, which not only shows a strong correlation between No. of followers and PageRank, but also proves its robustness in spite of domains of different quality.

3.4 Case Study

Domain 5 turns out to be a non-government organization. Many members of this domain belong to Guangdong Lions Club and show special interest in volunteer and giving. Table. 2 shows top 10 members (weibo ID) of this domain, from which we can see the rank by No. of Followers is almost the same as that by PageRank.

The first one, Weibo ID 1774929853, whose name is Jianye Qu, is a very active volunteer; The second user, ID 1915398804 , Li Cai, is the current President of the Lions' Club; While the third one, ID 2130085295, Siming Mai, is a former president. (Visit their profiles at http://weibo.com/id) Through detailed comparison, we can see that measuring domain influence by No. of followers or PageRank more accurately reflect users' real world profile and their social impact, which further confirms our motivation of introducing domains into influence analysis.

Domain 3 is an IT community, basically made up of programmers and other people in this industry. Table. 3 shows top 10 members by the three measures. Although a few members listed in Column Retweets is not that relevant due to noise caused by a larger community, we observe consistent conclusions as above where No. of Followers can nearly measure domain influence as well as PageRank. The first one, ID 1929644930, Shaoping Ma, is a famous professor on computer science in Tsinghua University; and this user turns out to rank No. 11 by No. of Retweets, which is not shown in the table below. The second, ID 2060750830, Dr. Hang Li, is the chief scientist in Huawei, former director in MSRA; And the third, ID 1355610915, Tao Jiang, is the founder of the popular Chinese technique blog –CSDN.

Table 2. Top 10 members of domain Lions' Club

By No. of Followers	By No. Retweets	By PageRank
1774929853 Jianye Qu	2056286021 Mingyan member	1774929853 Jianye Qu
1915398804 **Li Cai President**	2693203694 Jixian Yang member	1915398804 **Li Cai President**
2130085295 Siming Mai President	2001975055 Yanjun He TL	2130085295 Siming Mai President
1875775585 Gaosheng Cheng TL	1298998132 Luoye	1875775585 Gaosheng Cheng TL
1871086341 Aiyinshitan chairman	1788072593 Ziyun member	1871086341 Aiyinshitan chairman
1961172412 Yongzhong member	1615741212 Team of Life	1961172412 Yongzhong member
2006772367 Dong Zhou member	2534385342 Team of Guangda	2006772367 Dong Zhou member
2028775535 Dongmei secretary	2367431824 Lions' Poster	1843411063 Sixuan Li member
1471756485 Sicheng VP	1915398804 **Li Cai President**	1944360591 Changwei Huang
2189635834 Happy Lv member	2139637375 Team of Lingnan	2155791472 Zhaoxiang director

Table 3. Top 10 members of domain Computer Science

By No. of Followers	By No. Retweets	By PageRank
1929644930 Prof. Shaoping Ma	1682352065 Libo Zhou	1929644930 Prof. Shaoping Ma
2060750830 Dr. Hang Li	1705180884 Ou Chen CEO of Jumei	2060750830 Dr. Hang Li
1355610915 Tao Jiang CSDN	1896891963 Binxing Fang Beiyou Univ	1355610915 Tao Jiang CSDN
2098911447 Tieyan Liu MSRA senior	1182415487 unavailable	2098911447 Tieyan Liu MSRA senior
1936526225 Dr. Bin Wang ICT	1827652007 Prof. Jianrong Yu	1936526225 Dr. Bin Wang ICT
1715524730 Shen Jiang engineer	3196963860 Digital Network co.ltd	1715524730 Shen Jiang engineer
1614282004 Xueyong Cai Architecture	1419517335 Yuan Luo	3121700831 Prof. Zhihua Zhou
3121700831 Prof. Zhihua Zhou	2141100877 Visual Magazine	1614282004 Xueyong Cai Architecture
1918015782 Haifeng Wang Baidu	1670071920 YuzhuShi founder of Juren	1862459915 Kai Yu Baidu
1991303247 Laoshimu	3051172273 Yuguo Dai	1991303247 Laoshimu

4. CONCLUSION

This paper presents a systematic approach modeling domain influence in heterogeneous social networks. We utilize both structure and content information for spectral clustering and rank users respectively by No. Followers, No. Retweets and PageRank in each domain. Results show that correlation between No. of Followers and No. of Retweets has grown in domain influence than in general, especially when the domain is highly specialized, as shown in the case study of Domain 5 etc. Correlation between No. of followers and PageRank has also significantly grown with robustness in the sense of different domain qualities. To put it in another way, No. of Followers can nearly measure domain influence as well as PageRank, and No. of retweets is also more correlated to the other two measures, which is significant growth than previous work [4,5].

There are many potential future directions of this work. One interesting issue is to employ the classic Independent-cascade model in domain influence maximization; and another is to extend the present model to overlapped communities. Other relevant fields like viral marketing and recommender systems can also gain insights from this domain perspective.

5. ACKNOWLEDGMENTS

This study is partially supported by the National Natural Science Foundation of China (NSFC Grant No. 61272343), the Doctoral Fund of Ministry of Education of China (MOEC RFDP Grant No.20130001110032) as well as the "Undergraduate Research Training" program.

6. REFERENCES

[1] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In KDD '03, pages 137–146, 2003.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In KDD'08, pages 7–15, 2008.

[3] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In WWW '10, pages 601–610, 2010.

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In WWW '10, pages 591–600, 2010.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence on twitter: The million follower fallacy. In ICWSM '10, Washington, DC, 2010.

[6] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In KDD '09, pages 807–816, 2009.

[7] L. Liu, J. Tang, J. Han, M. Jiang, S. Yang. Mining topic-level influence in heterogeneous networks. In CIKM '10, pages 199-208, 2010.

[8] L. Tang, X. Wang, H. Liu. Community detection via heterogeneous interaction analysis. In DMKD '12, volume 25, issue 1 , pages 1-33, 2012.

[9] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, X. Li. Comparing twitter and traditional media using topic models. In ECIR '11, pages 338-349, 2011.

[10] W. Chen, Y. Song, H. Bai, C. Lin, E. Chang. Parallel spectral clustering in distributed systems. In TPAMI'11, volume 33, issue 3, pages 568-586, 2011