SocialLens: Searching and Browsing Communities by Content and Interaction

Hongyun Cai †, Vincent W. Zheng †, Penghe Chen †, Fanwei Zhu #, Kevin Chen-Chuan Chang ⋄, Zi Huang ‡

† Advanced Digital Sciences Center, Singapore # Zhejiang University City College, China

⋄ University of Illinois at Urbana-Champaign, USA ‡ School of ITEE, The University of Queensland, Australia

***Abstract*—Community analysis is an important task in graph mining. Most of the existing community studies are community detection, which aim to find the community membership for each user based on the user friendship links. However, membership alone, without a complete profile of what a community is and how it interacts with other communities, has limited applications. This motivates us to consider systematically profiling the communities and thereby developing useful community-level applications. In this paper, we introduce a novel concept of community profiling, upon which we build a SocialLens system[1] to enable searching and browsing communities by content and interaction. We deploy SocialLens on two social graphs: Twitter and DBLP. We demonstrate two useful applications of SocialLens, including interactive community visualization and profile-aware community ranking.**

Video: http://youtu.be/jieXE06Ki2Q

I. Introduction

Traditionally, the community-level analysis focuses on detection. Such community membership assists us to better understand the network structure. However, membership alone, without knowing *what a community is* and *how it interacts with others*, has only limited applications. For instance, we cannot rank communities by desired characteristics, or visualize community level interactions. In view of this critical lacking of community "understanding", we proposes systematic *community profiling*– to characterize the intrinsic nature and extrinsic behavior of a community– thereby enabling useful community-level applications. Fortunately, it is now feasible to profile communities, as richer user information is available online. Beyond traditional information (e.g., *friendship links*), there are more types of information, such as users' attributes, published content, and diffused content. While exploiting such rich user information has greatly improved recent community detection techniques [2], with the need to understand communities detected, it now provides us an unprecedented opportunity to achieve community profiling.

In this work, we introduce a new concept of **community profiling** [1]. A community profile should holistically characterize a community, both internally (i.e., what it is) and externally (i.e., how it interacts with others). For a community c, we define its *internal profile* as the distribution of X in c (i.e., $p(X|c)$), where X is a type of user information (e.g., content, attribute, etc). And its *external profile* is the probability of X being diffused between c and another community (i.e., $p(\text{diffusion}|X, c, c')$). In this work, we use *content* as X because it is the primary user information in many networks. E.g., in Twitter, users write tweets and also retweet from others; in DBLP, authors publish papers and cite papers. Therefore, the internal profile of community c is the content distribution of c, which characterizes what c is about; and the external profile is probabilities of certain content being diffused between c and other communities, which describes how c diffuses from others. Other types of X's may exist in different networks (e.g., attributes in Facebook). Thus, community profiling is a flexible concept. With the internal content profile and external diffusion profile, we enable searching and browsing communities by content and interaction. To demonstrate such usefulness, we build a **SocialLens** system. Next, we introduce its system architecture, including its input, output and applications.

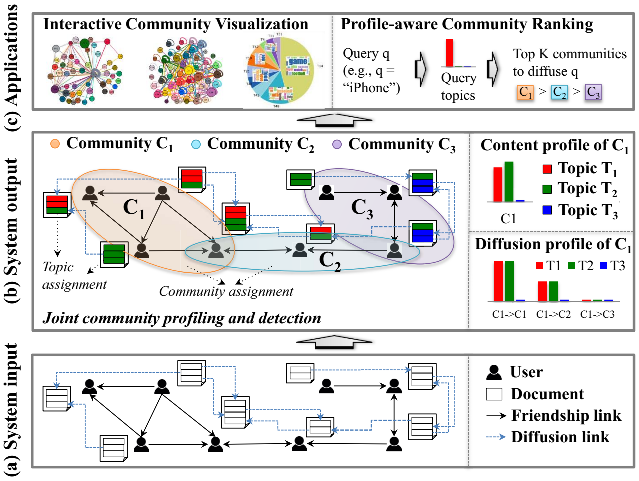


Fig. 1. SocialLens system architecture: input, output and applications.

II. System Overview

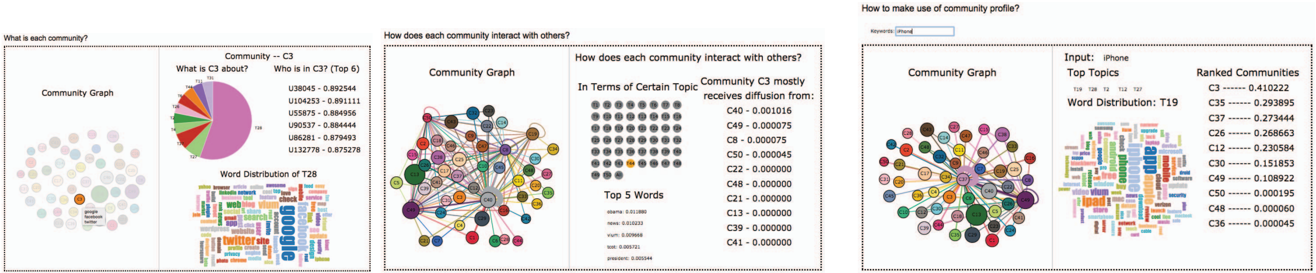
We illustrate SocialLens' system architecture in Fig. 1.

Input. As shown in Fig. 1(a), we have a network of users connected by friendship links. Each user generates documents and interacts with others by diffusion links. E.g., in DBLP, the authors are connected by co-author links. They publish papers and cite papers from others. In Twitter, the users are connected by followership links. They post tweets and retweet.

Output. As shown in Fig. 1(b), we detect a set of communities (e.g., c_1, c_2 and c_3) and infer the topic assignment for each document. For each community, we then infer a content profile

[1] http://sociallens.adsc.com.sg/.

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.194



U38045 - 0.892544
U104253 - 0.891111
U55875 - 0.884956
U90537 - 0.884444
U86281 - 0.879493
U132778 - 0.875278

obama: 0.011880
news: 0.010233
vium: 0.009668
tcot: 0.005721
president: 0.005544

C40 - 0.001016
C49 - 0.000075
C8 - 0.000075
C50 - 0.000045
C22 - 0.000000
C48 - 0.000000
C21 - 0.000000
C13 - 0.000000
C39 - 0.000000
C41 - 0.000000

Keywords: iPhone

Input: iPhone

C3 ------ 0.410222
C35 ------ 0.293895
C37 ------ 0.273444
C26 ------ 0.268663
C12 ------ 0.230584
C30 ------ 0.151853
C49 ------ 0.108922
C50 ------ 0.000195
C48 ------ 0.000060
C36 ------ 0.000045

(a) Interactive community visualization

(b) Profile-aware community ranking

Fig. 2. SocialLens demonstration scenarios.

(e.g., c_1 tends to publish on topics z_1 and z_2) and diffusion profile (e.g., c_1 tends to diffuse topic z_1 from itself and c_2).

Applications. As shown in Fig. 1(c), we enable two new applications with the learned community profiles:

- *Interactive community visualization*. We visualize how communities feature distinct contents and how they interact as in Fig. 2(a).In the left diagram of Fig. 2(a) (denoted as Fig. 2(a)-left for the simplicity of expression), we display 50 nodes, each as a community detected from Twitter. The size of each node indicates the size of a community. Users can click a specific node to see its content profile (the pie chart) and top users. E.g., C_3's top topic is T_{28}, which is about IT (see its tag cloud). In Fig. 2(a)-right, we plot how a community interact with others regarding to a topic (e.g., T_{44}) through a community graph. T_{44} is about politics as suggested by the top five words. We see that C_3 rarely diffuses about politics. Most politics diffusion happens within C_{40} (a politics-related community).
- *Profile-aware community ranking*. We can target communities, a larger-scale audience than individuals, for product/paper promotion. As shown in Fig. 2(b), people can input a query (one or more keywords) and find out which communities in Twitter are most likely to retweet about the query. E.g., given "iPhone" as a query, SocialLens first tries to understand what it is and finds that the most relevant topic is T_{19}, which is about "Apple", "app", etc. SocialLens then ranks the communities by their probabilities to retweet about the relevant topics of "iPhone" based on the learned community profiles. As expected, we find that C_3, an IT community, ranks the highest, followed by C_{35} which is a community discussing shopping.

III. Technical Novelty

Community detection focuses on identifying membership, whereas community profiling focuses on getting community content and diffusion profiles. A straightforward community profiling method is to first detect communities and then aggregate member users' content and diffusion in each community. Such an approach overlooks the fact that, as the members of a community are not only densely connected but also sharing same community profiles, detection can naturally leverage profiling. To close the loop between profiling and detection, we must solve them together.

It is not trivial to achieve joint profiling and detection. First, in terms of *model choice*, it is unclear how to model profiling and detection together. It is easy to first detect then profile communities, but it is not obvious how to utilize profiles in detection. To solve this, we take a profile-aware generative approach by introducing a set of community-level variables (membership, content profile and diffusion profile) to generate the observations. Second, there exists *data heterogeneity*– the input data, especially friendship and diffusion links, are heterogeneous, therefore we should tailor them to serve different needs. To solve this, friendship links are enforced to be denser within a community than across communities for community compactness, whereas the inter-community diffusion links are allowed to be denser on certain topics. Third, we need to ensure *model comprehensiveness*– user behaviors, especially their diffusion decisions, happen for many reasons rather than just community-level conformity. To solve this, we account for not only community diffusion profiles but also time-sensitive topic popularity and individual user preferences.

IV. Demonstration Scenario

We demonstrate SocialLens on two social graphs: Twitter and DBLP. For each graph, the system first loads a set of communities that are detected offline, and visualizes them as nodes in Fig. 2(a)-left. For interactive community visualization, users can click any community node to browse its top users, and content profile which is organized by topics in a pie chart with tag clouds. Besides, users can easily construct a community diffusion graph, based on either topic-aggregated or topic-specific diffusion strengths between any two communities, as in Fig. 2(a)-right. Users can also drag any community node to focus only on its diffusion with other communities. For profile-aware community ranking, users can query word(s) and find out which communities are most likely to diffuse (i.e., retweet/cite) that query. In Fig. 2(b), after users key in a query, SocialLens will return its top topics, as well as the community ranking based on the topics. Users can click any topic to see its tag cloud, and any community node to see its profile.

Acknowledgment

We thank the support of: National Natural Science Foundation of China (No. 61502418), Zhejiang Provincial Natural Science Foundation (No. LQ14F020002), Research Grant for Human-centered Cyber-physical Systems Programme at Advanced Digital Sciences Center from Singapore A*STAR, and NSF Grant IIS 16-19302. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the funding agencies.

References

[1] H. Cai, V. W. Zheng, F. Zhu, K. C.-C. Chang, and Z. Huang. From community detection to community profiling. *PVLDB*, 10(7), 2017.

[2] L. He, C.-T. Lu, J. Ma, J. Cao, L. Shen, and P. S. Yu. Joint community and structural hole spanner detection via harmonic modularity. In *KDD*, 2016.