*Sample** Diachronically Like-minded User Community Detection

Ali Fani
University of Windsor
afani@uwindsor.ca

Hossein Fani
University of Windsor
hfani@uwindsor.ca

ABSTRACT

This is a *sample* literature review for your proposal. The sections and subsections are *by no means* fixed and should be indeed changed or customized according to the proposal. For example, This paper provides a concise overview of the definitions, the underlying concepts, history and state of the art in user community detection and prediction.

KEYWORDS

More Specific, Specific, General, More General

1 USER COMMUNITY

The word community refers to a social context. People naturally tend to form groups, within their work environment, family, or friends. A community is a group of users who share similar interests, consume similar content or interact with each other more than other users in the network. Communities are either explicit or latent. Explicit communities are known in advance and users deliberately participate in managing explicit communities, i.e., users create, destroy, subscribe to, and unsubscribe from them. For instance, Google's social network platform, Google+[1], had Circles that allowed users to put different people in specific groups. In contrast, in this work, communities are meant to be latent. Members of *latent* communities do not tend to show explicit membership and their similarity of interest lies within their social interactions.

1.1 History

Probably the earliest account of research on community detection dates back to 1927. At the time, Stuart Rice studied the voting themes of people in small legislative bodies (less than 30 individuals). He looked for *blocs* based on the degree of agreement in casting votes within members of a group, called Index of Cohesion, and between any two distinct groups, named Index of Likeness [**?**]. Later, in 1941, Davis et al. [**?**] performed a social anthropological study on the social activities of a small city and surrounding county of Mississippi over 18 months. They introduced the concept of *caste* to the earlier studies of community stratification by social class. They showed that there is a system of colored caste which parsed a community through rigid social ranks. The general approach was to partition the nodes of a network into discrete subgroup positions (communities) according to some equivalence definition. Meantime, George Homans showed that social groups could be detected by reordering the rows and the columns of the matrix describing social ties until they form a block-diagonal shape [**?**]. This procedure is now standard and mainly addressed as blockmodel analysis in social network analysis.

[1]Google+ is no longer available since April 2019.

2 PRIOR WORK

In our work, we will focus on identifying and modeling the latent like-minded user communities detected in a given time period on online social networks. As a result, the main research area that is closely related to our proposal is community detection which we review in this section.

Existing community detection approaches can be broadly classified into two categories [2]; *link*-based and *content*-based approaches. Link-based approaches, also known as topology-based, see a social network as a paradigmatic example of a graph, whose nodes are users and edges indicate explicit user relationships. On the other hand, content-based approaches, also known as topic-based, mainly focus on the information content of the users in the social network to detect communities. The goal of content-based approaches is to detect communities formed around the topics extracted from users' information content. *Hybrid* approaches incorporate both topological and topical information to find more meaningful communities with higher quality. Recently, researchers have performed longitudinal studies on the task of community detection in which the social network is monitored in time intervals over a period of time [3**?** , 4]. The Time dimension results in a new *temporal* form of community detection which is the main motivation of our research. The following section includes the details of seminal works in each category. Herein, we use the terms 'graph' and 'network' interchangeably as well as the terms 'vertex', 'node', and 'user'.

2.1 Link Analysis

Link-based user community detection methods are primarily based on the homophily principle [**?**] where links between users are considered important clues for interest similarity and, as a result, densely connected groups of users imply a user community. In this line of work, the social network is modeled as a graph with nodes representing users and edges representing relationships or interactions. The primary principle considered in this line of work is *connectedness*, which means that connections within each community are dense and connections among different communities are relatively sparse. To this end, primitive graph structures such as components, cliques, k-plexes or other pseudo-clique structures are considered to represent user communities [2**?** **?**]. There are also graph partitioning (clustering) approaches which try to minimize the number of links between user communities so that the users inside one community have more intra-connections than inter-connections with other communities. Such approaches are based on iterative bisection: continuously dividing one group into two groups, while the number of communities which should be in a network is unknown. The Girvan–Newman approach [**?**] is one of the most commonly used methods in link-based user community

detection. It partitions the graph by gradually removing links with high betweenness centrality in a descending order. Betweenness centrality for a link is defined as the number of the shortest paths between any pairs of nodes that go through the link in a graph or network. A link with a high betweenness centrality score represents a bridge-like connector between two parts of a network such that the communication between many pairs of nodes through the shortest paths between them is affected by its removal.

Not all link-based methods perform well on large real-world networks that have many complex structural features such as sparsity, heavy tailed degree distributions and small diameters, among others. For empirical comparison of these algorithms in practice, see [? ?]. Nonetheless, link-based methods inherently fall short when the communities of interest need to take users' content similarity into account. This is mainly due to two reasons: *i*) there are many users on a social network that have similar interests but are not explicitly connected to each other; and, *ii*) explicit social connections do not necessarily indicate user interest similarity but could be owing to sociological processes such as conformity, aspiration, and sociability or other factors such as friendship and kinship that do not necessarily point to inter-user interest similarity [? ?]. There are also some special cases where link-based methods are not applicable like when the network is not available [?] or misleading, e.g., when links are fraudulent because of link-farmers (social capitalists) [?].

2.2 Content Analysis

With the development of social media, a significant amount of user-generated content, known as social content, is available within user networks. Users communicate and interact with each other in social network websites. Besides the links between users, huge amounts of textual content are generated as well. Along with rich information in social network structure, user graphs can be extended with textual information on nodes. In social networking sites, users maintain profile pages, write comments and share articles. In photo and video sharing sites, users use short texts to tag photos and videos. In microblogging websites, users post their status updates. Therefore, researchers have explored the possibility of utilizing the topical similarity of social content. They have been proposing topic-based community detection methods, irrespective of the social network structure, to build like-minded communities of users [5? ? ? ? , 6].

Most of these content-based methods have been inspired by latent Dirichlet allocation (LDA) [1] in one way or another and focused on probabilistic generative models based on textual content [5?]. For example, Zhou et al. [?] have modeled communities based on topics of interest through a community-user-topic generative process to identify user communities. In their work, communities follow multinomial distribution over topics with Dirichlet priors where each user is posting about her topics of interest based on the conditional probability of a topic given each community. Abdelbary et al. [?] have identified users' topics of interest and extracted latent communities based on the topics utilizing Gaussian Restricted Boltzmann Machines. Yin et al. [6] have integrated community detection with topic modeling in a unified generative model to detect communities of users who are coherent in both structural relationships and latent topics. In their framework, a community can be formed around multiple topics and a topic can be shared between multiple communities. Sachan et al. [5] have proposed probabilistic schemes that incorporate users' posts, social connections and interaction types to discover latent user communities in social networks. They have considered three types of interactions: a conventional tweet, a reply tweet and a re-tweet. Author-Topic-Community model [?], Author-Topic model [?] and Community-User-Topic model [?] are other variations of latent Dirichlet allocation (LDA), which are also proposed to identify user communities.

2.3 Temporal Analysis

All the above methods do not incorporate *i*) temporal aspects of users' topics of interests, and/or *ii*) dynamics of social links and undermine the fact that users of communities would ideally show similar contribution or interest patterns for similar topics and/or similar network neighborhood evolution throughout time. As a matter of fact, many content based and link based methods assume that the structure of the network and the topics discussed by the users remain stable over time, which can be a limiting assumption in practice. While a myriad of work has addressed the dynamics of social links in user community detection, i.e., dynamic community detection [? ?], there have been a few works that have considered temporal aspects of users' topics of interests as an explicit dimension when identifying user communities in social networks [3? , 4].

The work by Hu et al. [3] is among the pioneers to consider temporality in user-generated contents through a generative process, which models how users and topics are related to each other and co-evolve over time. Their model, namely Group Specific Topic over Time (GrosToT), learns a specific time-aware probability distribution known as the community-topic-time distribution addressing how communities and topics are associated with each other over time. Assuming the number of topics K and the number of communities C are known in advance, the model associates Dirichlet distributions for topics over words, communities over users, and topics over communities with different parameters, respectively. Also, a Dirichlet distribution for time is assigned given topic-community pairs. A user is a member of a community according to the assigned community-user distribution. As seen, the model is based on the idea that there is a tight interrelation between communities and topics. This prevents integration of other topic detection methods for the task of community detection.

In our research, we follow the same underlying hypothesis related to topics and temporality as by Hu et al. and Liang et al., i.e., the evolution of user-generated content is dynamic over time (temporal) and users' interests can evolve over different time intervals. In contrast to dynamic community detection methods, however, we assume that the social network structure is static and remains stable over time. The main reason for this assumption is that the social network structure has a significantly lower pace of change compared to how fast content is generated over time and distributed across the online social network [?].

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[2] Santo Fortunato. 2010. Community detection in graphs. *Physics Reports* 486, 3/5 (2010), 75 – 174.
[3] Zhiting Hu, Junjie Yao, and Bin Cui. [n.d.]. User Group Oriented Temporal Dynamics Exploration. In *The 28th AAAI Conference on Artificial Intelligence, 2014.*
[4] Zhiting Hu, Junjie Yao, Bin Cui, and Eric P. Xing. 2015. Community Level Diffusion Extraction. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015.* 1555–1569.
[5] Mrinmaya Sachan, Danish Contractor, Tanveer A. Faruquie, and L. Venkata Subramaniam. [n.d.]. Using content and interactions for discovering communities in social networks. In *The 21st World Wide Web Conference 2012, WWW 2012.*
[6] Zhijun Yin, Liangliang Cao, Quanquan Gu, and Jiawei Han. 2012. Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling. *ACM TIST* 3, 4 (2012), 63:1–63:21. https://doi.org/10.1145/2337542.2337548