Finding Structure in Blogs: Bipartite Networks Analysis

[Invited Presentation, Extended Abstract] * †

Bosiljka Tadić‡
Department of Theoretical Physics
Jožef Stefan Institute
1000 Ljubljana, Slovenia
bosiljka.tadic@ijs.si

Marija Mitrović§
Department of Theoretical Physics
Jožef Stefan Institute
1000 Ljubljana, Slovenia
marija.mitrovic@ijs.si

ABSTRACT

Temporal patterns of activity on Blogs (posting, reading, commenting, comment-on-comment) contain valuable information about user behavior, which leads to potentially new type of social clustering in the Blog space. Here we show how the structure in Blog space can be retrieved from the data by mapping onto a bipartite graph and using the appropriate methods of complex networks, including the spectral analysis of graphs [4, 3]. With the analysis of (almost) complete set of data from B92 Blogsite since its opening, we demonstrate how the user communities emerge in time and what are possible underlying mechanisms of this structure.

Categories and Subject Descriptors

E.m [**Data**]: Miscellaneous; I.2.4 [**Computing Methodologies**]: Artificial Intelligence—*Knowledge Representation Formalisms and Methods*

Keywords

Complex networks, Blog structure, Cyber communities

1. MOTIVATION

In recent review *The convergence of social and technological networks* [1] Kleinberg stressed that the "Internet-based data on human interaction connects scientific inquiry like never before". Underlying the on-line social interaction facilitated by the communication technology (e.g., on Blogsites, Facebook, MySpace, Wikipedia, Digg, and other), "there is a broader process at work, a growing pattern of movement through online spaces to form connections with others, build virtual communities, and engage in self-expression"[1]. The social clustering emerging through these interactions is both new and fast developing and playing an important role in everyday life of modern society. In the classical approach, social interactions have been studied by mapping onto complex networks of connected individuals, who are forming groups with traditional social meaning. Grouping in the cyber space, however, is different and not necessarily related to conventional friendship, family or business relationships. This implies other mechanisms which drive people's activity in the cyber space. Recent studies [5] indicate diversity in people's interests to posts, that might be related with *quality, emotional, or moral contents of the posted material* and its interference with users preferences and personal profiles. Consequently, new scientific methods are needed for the analysis of cyber communities. In our approach, we map large datasets about users and their posts and comments collected from Blogs onto suitably defined bipartite graphs, in which both users and their posts play an equal role. We then analyze user communities [3, 4] and identify posts which cause their clustering in the Blog space.

2. DATA STRUCTURE

We consider large datasets collected from different Blog sites over extended period of time (from few weeks to few years). While the Blog sites can differ in internal organization and history containing a different information about bloggers, post and comments, the common feature of all blogs which enables us to study social interaction network is the existence of unique identification of users (every blogger has to be registered under unique ID). Further details which affect the analysis may vary depending on the blogsite. For instance, bloggers are allowed to write posts (B92 Blog), select a news story (Digg), or just leave a comment on posts written by professionals (BBC Blog). Regarding the subject categories, most Blogsites have in-advance determined categories of posts, the exception is the B92 Blog where users write their story without obligation to adhere with a predefined category. Consequently, the internal structure with user communities on B92 posts emerges in a self-organized manner through user interactions on posts and comment-

†**Work suported** by FP7 project CYBEREMOTIONS and P1-0044 (Slovenia).

‡Presenting author.

§On leave from the Scientific Computing Laboratory, Institute of Physics, Belgrade, Serbia.

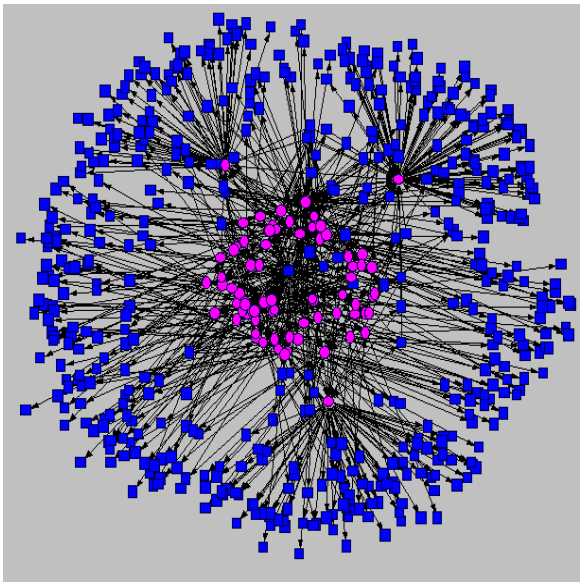

Figure 1: An example of directed bipartite network related to one specific post from Blog B92 data. Post and comments are represented by (blue) squares, while user nodes are marked by (pink) circles.

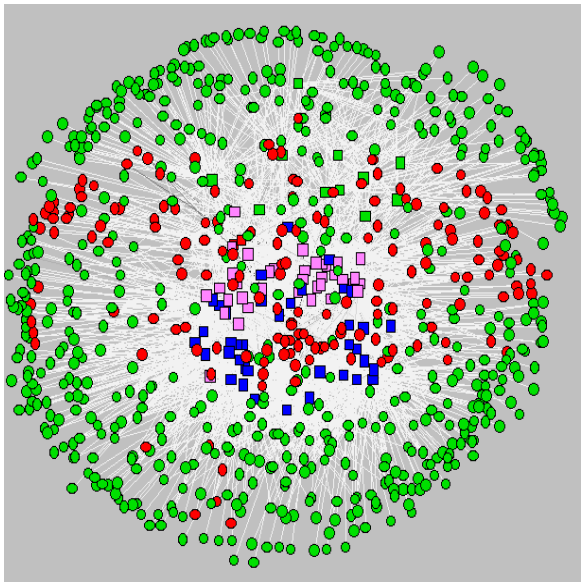

Figure 2: Weighted bipartite network of popular posts (□) and linked users (○) on B92 Blog. Different colors indicate community structure, as found using the weighted maximum likelihood method [2].

on-comment activity. On the other hand, the availability of B92 posts is time limited to seven days while this kind of limitation is not present in other Blogsites analyzed. Digg and B92 Blogs data also contain information about the ID of a comment which is commented, allowing a network of links with more intricate structure.

3. BIPARTITE NETWORKS OF BLOGS

Blogs are mapped onto bipartite network with *users* representing one partition, and *posts and comments*, as the other partition. By definition, in bipartite networks no direct link occurs between nodes within the same partition, thus two users in our network interact with each other only trough posts and related comments. In the data we have $i_U = 1, \ldots, N_U$ users and $i_B = 1, \ldots, N_B$ post nodes ($N_U + N_B \approx 500000$). The links between users and posts are defined according to their action, as follows: If the user i_U is the author of the post j_B or has left the comment $kcmj_B$ on that post, a directed link is inserted pointing from $i_U \rightarrow j_B$, and similarly from $i_U \rightarrow kcmj_B$. Each post (comment) has only one author and thus one incoming link. Whereas, multiple outgoing links may occur, pointing from a post towards the users who commented it, i.g., $l_B \rightarrow p_U$, if user p_U commented post l_B. In the case (B92) where the comment-on-comments is allowed, an outgoing link occurs also from the post $l_B \rightarrow m_U$ when the user m_U commented a comment, say $rcml_B$ related to that post. In this case the link is also created from the comment $rcml_B$ to the user m_U. As an example, we show a part of such network in Fig. 1. Note that in the mapping multiple links occur, e.g., if the user replied to comments on his post or comments. The network representation of the Blog data allows us to study quantitatively the posts-mediated interactions between users and how the behavior of users affects the structure of the Blogs. More details of the methodology and results are given in Ref. [3].

3.1 Structure in weighted bipartite networks

Through a systematic analysis of Blog networks, we found that the popular posts (posts with many comments) fall in a separate category, that can be suitable represented by *weighted bipartite network* [3]. The weight of a link, W_{ij} between user i and post j is built through multiple comments (replies to comments) on popular posts. Consequently, in addition to the topology of links, the weights of links also affect the community structure on such networks, which can be detected by the spectral analysis [4] and maximum-likelihood methods of weighted graphs [2]. Example of the community structure of popular posts is shown in Fig. 2

In conclusions, the bipartite network representation of Blog data with the appropriate graph theory methods, on one side, and fine granularity of the data, on the other, provide a powerful methodology to study subject-oriented communities in the cyber space.

4. REFERENCES

[1] J. Kleinberg. The Convergence of Social and technological Networks. *Communications of the ACM*, 51, 2008.

[2] M. Mitrović and B. Tadić. Search of Weighted Subgraphs on Complex Networks with Maximum Likelihood Methods. *Lecture Notes in Computer Science*, 5102:551–558, 2008.

[3] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in blog space. *Europ. Phys. J. B*, 2009.

[4] M. Mitrović and B. Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Physical Review E*, 80, 2009.

[5] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 2009.