Trends in collaborative networks

Project proposal

By Alexey Reznick 306483371

Advisor Dr. Osnat Mokryn

Contents

Abstract

This is a proposal for a research project in the area of Complex Network analysis. The study of complex networks is an active area of scientific research inspired largely by the empirical study of real-world networks such as computer networks and social networks (1), which are samples of complex networks. The proposed project will focus on a subset of real-life social networks termed "Collaborative Networks" – networks that are defined by collaboration of people or organizations. This proposal outlines the field that we're planning to study and some possible research directions.

Introduction

Various real-life networks, such as social; biological; and technological, are considered complex, since they display substantial non-trivial topological features. Patterns of connection between the nodes of those networks are neither purely regular nor purely random. Such features include a heavy tail in the degree distribution, a high clustering coefficient, community structure, and hierarchical structure. In contrast, many of the mathematical models of networks that have been studied in the past, such as lattices and random graphs, do not show these features. (1)

We describe here well-known subsets of complex networks, and briefly describe the features that characterize these networks.

Definitions

Throughout this paper, the following terms will be used interchangeably:

- **Network/Graph** – Since every network can be represented by a graph, we will use both terms in an interchangeable manner. The term "graph" is used to describe a graphical representation of the network, or in context in which graph theory terms or algorithms are discussed.
- **Node/entity/actor** and **edge/link/relation** – Those terms describe a node and an edge on a graph respectively. Since collaborative networks connect people or organizations, nodes represent either entities or actors and edges represent either links or relations (The convention used in sociology)

Scale-Free networks

A network is termed scale-free (2) if its degree distribution, i.e., the probability that a node selected uniformly at random has a certain number of edges (degree), follows a long tail distribution. Example for a common long tail distribution in complex networks is a power law. In a power law distribution the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $P(k) \sim ck^{-\gamma}$ where c is a normalization constant and γ is a parameter whose value is typically in the range $2 < \gamma < 3$, although occasionally it may lie outside these bounds.

The power law implies that the degree distribution of these networks has no characteristic scale.

Small-world networks

A network is termed a small-world network if while most nodes of the network are not neighbors of each another, they can be still reached from each other by a small number of hops. Specifically, a small-world network is defined to be a network where the typical distance *L* between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes *N* in the network, that is $L \propto \log N$ (3).

Cohen and Havlin showed analytically that scale-free networks are ultra-small worlds. In this case, due to hubs, the shortest paths become significantly smaller and scale as $L \propto \log \log N$ (4)

Additionally, small world networks, first discovered by Watts and Strogats, were shown to have a high clustering coefficient. That is, the probability of two nodes to have a connecting link between them is higher if they share a common neighbor.

Modeling complex networks

Few generative models for complex networks were proposed over time. The mostly widely known generative model for a subset of scale-free networks is Barabási and Albert's (5) rich-get-richer model in which each new node creates an edge to existing nodes with a probability distribution which is not uniform, but proportional to the current in-degree of nodes. This model, termed preferential attachment model, was originally proposed by Derek J. de Solla Price in 1965 under the term cumulative advantage, but did not reach popularity until B-A rediscovered it under its current name (BA Model). According to this process, the probability of a node to attract more incoming edges is proportional to its degree. Hence high degree nodes have a higher probability to increase their degree, and hence the term rich get richer. This generates a power-law for node's in-degrees. The BA model was created to model the WWW, but the resulting graph differs from the actual Web graph in properties such as the presence of small tightly connected communities.

Collaborative Networks

Collaborative (collaboration) network is a network that consists of nodes that represent entities (organizations, people) and edges that represent some sort of collaboration between them. For example: a collaborative network between mathematicians, where an edge represents a paper written together. This network is famous by assigning each mathematician an "Erdős number" – the "collaborative distance" between the mathematician and Paul Erdős, who is one of the most prolific modern writers of mathematical papers (6). Another example of a similar collaborative network is a network of movie actors which are linked by co-acting in the same movie. Similarly, this network is known for assigning each movie actor a "Bacon number", a collaborative distance between the actor and Kevin Bacon.

Trends in collaborative networks

Another important research field is the temporal patterns describing network's evolvement and evolution. Those patterns give us valuable insights into the behavior of the real-life phenomena, which is modeled by the network. Different networks exhibit different sets of temporal patterns, ranging from circadian and weekly rhythms to seemingly random bursts of activity.

Since collaborative networks model sociological phenomena, we term such bursts in activity as trends. Observing a trend in a collaborative network usually indicates that an aspect of collaboration has become popular among the subject population. In case of one-time and long-term collaboration, it would mean that the entity has gained popularity and in case of crowd-collaboration, it would be the project, which the entities collaborate upon. Popularity of an entity means that it attracts more incoming links, than it would, based on preferential attachment only.

We think that studying such trends could yield deeper understanding of the behavioral patterns of subject population.

Wikipedia as a collaboration network

Wikipedia is a crowd-sourcing project, which aims to collaboratively create a free encyclopedia, in multiple languages. It uses Wiki software, a type of collaborative software that was invented first by Ward Cunningham in 1995 (7). He created a simple tool for knowledge management and decided to name it "Wiki Wiki Web" using the Hawaiian term "wiki" for "quick" and with allusion to the WWW. Briefly a Wiki is a collection of hypertext documents that can directly be edited by anyone. Every edit is recorded and thus can be retraced by any other user. Each version of a document is available in its revision history and can be compared to other versions.

According to (8) (9) currently the English Wikipedia has 28,333,283 pages, from which 4,067,075 are content pages (pages can be content-pages, talk-pages, redirects, etc.) The English Wikipedia has 17,591,568 of registered users, and there are over 3 million edits per month since March 2006. The total number of edits to date is 559,420,488.

Our work: Analyzing trends in collaboration networks

We define few types of collaborative networks, characterized by the type of collaboration that forms the links in the network.

The first type is a **one-time collaboration network**, such as co-authoring a scientific paper, or co-acting in a movie. This type of network (graph) is undirected, and can have more than one edge between nodes in case that the entities in question have multiple collaborations (it can also be represented as edges that have weights, which correspond to the number of collaborations between the actors).

The second type is a **long-term collaboration network** – a mutual state in which unites the entities. A good example for such collaborative network would be a "social network" in which the nodes are the people and the edges that connect them represent friendship bonds between them. This type of

network creates an undirected (since friendship is usually mutual) graph with only a single edge possible between two nodes. Since the mutual state (friendship) is not limited in time, whenever it ends – the link between the nodes is removed.

The third type of collaboration network is a **crowd-collaboration network**. In this type of network entities (people, organizations) collaborate on a project of some sort, without being directly connected. Examples of such networks are software developers collaborating on different open-source projects, various crowd-sourcing projects and wikis, such as Wikipedia in which editors collaborate on different articles, crowd-funding projects in which users can fund together projects that they like, and many more.

In this work we focus on the third type of collaboration network, as it was defined previously – the crowd-collaboration networks, and specifically on the Wikipedia crowd-sourcing project. We intend to examine the temporal evolvement of the network, and try to distinguish patterns that describe the collaborations of Wikipedians (users that edit Wikipedia) on articles. We further plan to analyze the effect that bursts of user activity on an article or a group of articles (trends) has on the network's structure, and try to establish some underlining properties that define those trends.

Previous work: Trends in Wikipedia

Wikipedia is a crowd-sourcing project, and it's affected by the changes of interests of its users. Moreover, some researchers suggested that a significant amount of activity performed by the Wikipedia contributors is motivated by news and mass media (10). Therefore it is reasonable to expect that increased editorial activity of particular topics may reflect extreme events in nature and society.

However, previous work (11) that have studied the statistical properties of edit patterns in Wikipedia, have discovered only one pattern that characterize a stream of edit events (edit-stream): local clustering of edit events. This pattern is caused by a combination of two factors: intermediate saves performed by the same author and edits that are immediately re-edited by other author (edit-wars). Other than that, no long-term correlations were discovered in the edit-stream. This hints that editorial activity is not driven by the shifts in popularity of topics of general interest in society.

Although this feels counterintuitive, the apparent randomness of edit events can be caused by superimposition of few factors:

- The editing process is controlled by a collaborative network of editors, which are dedicated to improving the content of Wikipedia, and coordinate their efforts via discussion pages and to-do lists.
- External factors (for example a strike of inspiration of one of the editors)
- Continuous process, of few uncoordinated editors that feel responsible for one or more topics, and constantly updating them.

Bibliography

1. Complex Network. *Wikipedia.* [Online] [Cited: 10 2, 2012.] http://en.wikipedia.org/wiki/Complex_network.

2. **Barabási, Albert-László and Bonabeau, Eric.** Scale-Free Networks. *Scientific American Magazine.* May 2003, pp. 50-59.

3. **Watts, Duncan and Strogatz, Steven .** Collective dynamics of 'small-world' networks. *Nature.* 1998, Vol. 393, pp. 440-442.

4. **Cohen, Reuven and Havlin, Shlomo.** Scale-Free Networks Are Ultrasmall. *Physical Review Letters.* February 4, 2003, Vol. 90, 5.

5. **Barabási, Albert-László and Réka, Albert.** Emergence of Scaling in Random Networks. October 15, 1999, Vol. 286, 5439, pp. 509-512.

6. **Newman, M. E. J.** The structure of scientific collaboration networks. *Proceedings of the National Academy of the United States of America.* January 16, 2001, Vol. 98, 2, pp. 404-049.

7. **Leuf, Bo and Cunningham, Ward.** *The Wiki Way: Quick Collaboration on the Web.* s.l. : Addison-Wesley Professional, 2001. ISBN-13: 978-0201714999.

8. Wikipedia Statistics English. *Wikipedia Statistics.* [Online] [Cited: October 3, 2012.] http://en.wikipedia.org/wiki/Special:Statistics.

9. *Measuring Wikipedia.* **Voß, Jacob.** 2005.

10. *Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource.* **Lih, A.** Austin, Texas : s.n., 2005. 5th International Symposium of Online Journalism.

11. *Fluctuations in Wikipedia access-rate and edit-event data.* **Kämpf, Mirko, et al., et al.** 391, Physica A, pp. 6101-6111.