Exploring the Micro-Structure of Social Graphs

Supervision

Arnaud Legout, (main supervisor), Inria, Researcher (HDR), DIANA team.
Frédéric Giroire, CNRS Researcher, I3S/COMRED/COATI.

Scientific themes

Graph analysis, social graphs

Context

Twitter is the most popular micro-blogging service in the world. It allows its users to exchange short messages (tweets) that are limited to 140 characters. It was created to enable people to find out what is currently happening with people and organizations they are interested in. The relation between users on Twitter is different from classical social networks like Facebook. Instead of bidirectional friendship links that are initiated by one user and accepted by another, Twitter uses the concept of following. Users can follow other users they are interested in, which means they subscribe to all the messages they sent. So, the links on Twitter are unidirectional, if someone follows you, you don't need to follow back. Twitter is a very interesting object of study because the unidirectional model of relationship can represent more schemas of real-life communications, thus it has a huge societal impact.

Directed social networks are growing in importance because they are the main source of information for many people, but also because the information exchanged in these networks feed traditional media (such as TV or newspaper). Also, social networks are growing in size: end of 2014, Twitter had more than 1 billion accounts.

However, little is known about the inner structure of the directed social graphs. Indeed, high level statistics such as node degree, diameter, number of triangles, etc. are interesting, but do not give any information on the inner properties of the graph. Saying that such graphs are small words is also most of the time wrong even at a first level approximation. We exhibited a macrostructure of the Twitter social graph [p1], but in this macrostructure we have the largest strongly connected component gathering together 50% of all accounts. Exploring the properties of this component that represents the regular Twitter activity is important but complex because there is no appropriate theoretical tool to analyze this component.

Objective

In this PhD, we want to focus on the analysis of directed social graphs. In particular, we want to invent new graph analysis tools and metrics exhibiting specific social properties of the graph. Indeed, the biggest challenge is not to find a new network metric, but to link this metric to a social property.

This thesis will be conducted along three complementary axes:

- A measurement activity whose goal is to collect, process, and analyze the social graph data.
- A graph theoretical analysis of the collected data.

- An experimental validation of the link between the identified graph properties and the social properties.

1. Measurement activity

We collected in July 2012 the full Twitter social graph [p1]. However, the graph is evolving fast, and even if many studies can be performed on the 2012 graph, all results must be validated with the latest structure of the graph. In order to make such a validation, the student will need to perform new measurements using, for instance, sampling and distributed measurements. The NEPI [3] framework enables the deployment of large-scale measurements on unreliable distributed platforms such as Planetlab.

2. Graph analysis

Designing new theoretical graph analysis tools that can be linked to social properties is a complex task because it requires understanding the physical properties behind the graph properties. So the graph cannot be considered as an abstraction, but must be considered in the context of the social profile of each node. This makes the number of parameters to take into account in the modeling much broader. Eventually, we want to link new graph properties to social properties. For instance, Backstrom and Kleinberg [1] show that dispersion is a much better property than embeddedness [2] to exhibit a social tie between romantic partners.

3. Experimental validation

The validation of a new graph property is difficult because it must be linked to a social property that might not appear in a straightforward way in analyzed graph. For this reason, we might need to perform correlations with other social networks (for instance, Facebook or Wikipedia) or to experiment with profiles.

More specifically, we will investigate the following points, which may serve as starting points for the work of the Ph.D. candidate.

- The dispersion metric was defined on the Facebook graph to identify romantic partners among a user's friends. However whereas the Facebook graph is undirected, the Twitter graph is directed. This means that metrics must be adapted to a directed graph. Is it possible to define metrics such as the dispersion on a directed graph? Can we define new metrics to identify social links among users on directed graphs?

- Users might use Twitter differently depending on countries or social habits. Can we distinguish differences on structures of the subgraphs corresponding to users of different countries? We will first extract these subgraphs and test their proximity/remoteness for different metrics. We will then explore if the differences (if any) corresponds to different social meanings or to different uses of the social media.

- Twitter is a key source of information for traditional media such as TV or newspapers. It is also becoming a privileged media for public figures like politicians. We will explore whether these important actors correspond to specific subgraph structures. More generally, we will study the graph structures induced by the different categories of users and the impact of these subgraphs on the global graph.

- These investigations on the social meaning of the relations on Twitter may answer different unexplained phenomena observed in the social graph [p1]. As an example, there are several anomalies on the distribution of the followings and followers. Some parts of the distribution are very irregular. It would be very interesting to see if these irregularities can be explained by the specific profiles of particular users.

Each of these points will be treated according to the three axes of the Ph.D., extracting some adequate subgraphs, analyzing it, and then validating the conclusions on additional data.

Complementary of teams involved in the project

This project is supported by both DIANA and COATI. Each team brings a unique set of competencies that make the subject of this Ph.D. thesis both viable and achievable.

DIANA is a team conducting research in the domain of networking, with an emphasis on the design, implementation, and analysis of new network architectures with a special focus on Internet users rights.
The key expertise of the DIANA team on large scale measurements and experiments and in particular its expertise on Twitter [p1, p2] allows to tackle all the measurement and experimental part of this project.

COATI (Combinatorics, Optimization, and Algorithms for TelecommunIcations) is a joint project-team between INRIA Sophia Antipolis – Méditerranée and the I3S laboratory, with a strong expertise in algorithms, graph theory and combinatorial optimization applied to telecommunication (wired and wireless) networks and, in particular, the study of graph properties in order to design efficient algorithms [p3] and the study of algorithms for large systems [p4]. This is a key expertise to tackle the algorithmic challenges of the proposal.

Relevance with the Labex axes and priorities

This project is relevant for the axes “Stockage, transfert et traitement de grands volumes de donné” and “Interaction avec les SHS”.

“Stockage, transfert et traitement de grands volumes de données”
The Twitter social graph represents today around 1 billion vertices and tens of billions of edges. This makes this social graph a very large graph. The challenge here is not in the storage or transfer of this data (that is in the order of the terabyte, so not very large in absolute size), but in the processing of this graph with complex metrics. Indeed, some simple graph computations scale to virtually an unlimited number of vertices (e.g., node degree), but others have a complexity that is polynomial or even exponential. We want to design metrics that scale to large graphs and to be able to perform sophisticated analysis on large realistic graphs.

“Interaction avec les SHS”
Analyzing the twitter social graph is a very good way to understand the way social links are created among human beings. Our analysis, such as exploring romantic partners, is very relevant in the context of SHS because it exploits an existing social graph to discover social relationships among human beings. Moreover, such an analysis raises also privacy concerns and is therefore relevant for the axis “Sécurité, respect de la vie privée et neutralité du réseau”

Selected publications

[p1] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. **Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph**. In *Proc. of ACM SIGMETRICS'14*, June 16--20, 2014, Austin, Texas, USA.
[p2] Giovanni Neglia, Xiuhui Ye, Maksym Gabielkov, and Arnaud Legout. **How to Network in Online Social Networks**. In *Proc. of IEEE NetSciCom'14*, May 2, 2014, Toronto, Canada.

[p3] Fomin, F. V., Giroire, F., Jean-Marie, A., Mazauric, D., & Nisse, N. (2014). **To satisfy impatient web surfers is hard**. *Theoretical Computer Science*, 526, 1-17.
[p4] Caron, S., Giroire, F., Mazauric, D., Monteiro, J., & Pérennes, S. (2013). **P2P storage systems: Study of different placement policies**. *Peer-to-Peer Networking and Applications*, 1-17.

References

[1] Lars Backstrom and Jon Kleinberg. 2014 **Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on Facebook**. In *Proc. of the 17th ACM conference on Computer supported cooperative work & social computing* (CSCW '14). ACM, New York, NY, USA, 831-841.
[2] Marsden, P. V., and Campbell, K. E. **Measuring tie strength.** Social Forces 63, 3 (Dec.1984), 482-501
[3] Alina Quereilhac, Mathieu Lacage, Claudio Freire, Thierry Turletti and Walid Dabbous. **NEPI: An integration framework for Network Experimentation**. In *Proc. of the 19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2011.*