Decentralized Network Analysis: a Proposal

Alberto Montresor
University of Trento
Dipartimento di Ingegneria e Scienza dell'Informazione
via Sommarive 14, 38100 Trento, Italy
montresor@disi.unitn.it

Abstract

In recent years, the peer-to-peer paradigm has gained momentum in several application areas: file-sharing and VoIP applications have been able to attract millions of end users, while large-scale distributed computing frameworks, including the Grid, have proven their ability of attacking large scientific problems. We believe, however, that the potential of the P2P approach has not been completely exploited yet. The goal of this position paper is to propose another scientific area where the P2P cooperation paradigm could be profitably adopted: network analysis*, i.e. the mathematical characterization of the main graph-theoretic properties of a large-scale network. We discuss the potential issues that must be confronted with when a decentralized approach to network analysis is taken, and we propose a preliminary research plan.*

1 Introduction

Network is a heavily overloaded term used to describe physical artifacts like electrical circuits, transportation systems, and communication networks, as well as more "virtual" phenomena like social networks, food webs, and protein networks. *Network analysis* refers to the analysis of (potentially large) networks through graph theory, with the purpose of identifying their structural properties and features. While social sciences have a long tradition in this field, it is only in recent years that network analysis has emerged as a multi-disciplinary paradigm for the study of complex systems in areas as diverse as computer science, physics, epidemiology, biology, bibliometrics, etc.

A large collection of theoretical definitions have been identified to describe a multitude of network properties, each of them suitable for a particular task. We provide here a few examples, aiming at illustrating the broad applicability of these concepts.

- *Element-level analysis* formalizes the intuitive feeling that some network elements (vertexes or edges) are more important (central) than others. Well-known examples are: (i) *betweenness centrality*, that measures the number of shortest paths traversing a vertex; in a communication network, it may be used to evaluate the "stress" that such vertex has to sustain; (ii) Google's PAGERANK (a variant of *eigenvector centrality*), that sorts web pages based on their importance.
- *Group-level analysis* is aimed at finding groups of elements, for example by identifying strong linkages among its members. *Connectivity* and *subgraph properties*, as well as *clustering* algorithms are covered by this topic.
- *Network-level analysis* describes essential properties of an entire system, with the aim of differentiating between distinct classes of networks and exposing global information about the network. Such information can be used to optimize the behavior of other algorithms that operate over the network. Well-known examples are: (i) *average path length* that quantify the breadth of a network and help in tuning communication parameters such as time-to-live (TTL); (ii) *robustness*, that measures the ability of a network to withstand random and coordinated attacks, and can influence the replication degree needed to obtain a specified availability level.

The design of efficient algorithms for the computation of such properties over large networks is an active area of research. Almost all of the proposed algorithms are based on a completely natural, but very strong assumption: data describing the network to be analyzed are concentrated in a single location, where one or more computing units operate on them based on a "global view" of the entire network.

While this assumption simplifies the design of such algorithms, it has very important implications:

- if the network description is not already available at

978-0-7695-3315-5/08 $25.00 © 2008 IEEE
DOI 10.1109/WETICE.2008.36

a central location, a potentially large amount of data must be transferred;

- the maximum size of networks that can be analyzed is limited by the computational and storage power of the centralized analysis unit;

- only off-line analysis is possible;

- data owners must be willing, and in some cases even legally authorized, to transfer their data to third parties for analysis.

As an instance of such problems, consider the huge amount of information that telecom operators hold about their clients; performing network analysis on such data in a traditional, centralized way would be rather difficult, both from a technological and a legal point of view.

While the examples proposed so far are limited to technological processes, an incredible opportunity could come from the analysis of the large-scale social, biological and economic networks.

2 Decentralized Network Analysis

We believe that it would be possible, and useful, to go beyond the the centralization assumption, and design algorithmic techniques for the decentralized analysis of large-scale networks.

Decentralization means that a distributed collection of machines cooperate to evaluate network-wide properties without each single node having access to a global, complete view of the analyzed network.

Several advantages would derive from such a decentralized approach: larger problem instances could be attacked and solved, thanks to the combined computational power of multiple machines; access to expensive computing facilities would not be required any more; on-line analysis would be possible, allowing participating nodes to promptly react to the result of such analysis; decentralized agents could be executed by the owners of data, enabling the communication of aggregated information without requiring neither large data transfers nor the communication of valuable data to third parties.

The scientific community has not identified the decentralized analysis of large-scale networks as an independent research topic yet. While parallel algorithms already exist, they are often limited to multi-processors and multi-core systems; only an handful of algorithms for the distributed computation of specific properties exist (e.g., betweenness centrality, eigenvector centrality, clustering, etc.); but a coherent vision of the field is still missing.

3 A Research Plan

The position of this paper is that this void should be filled and such coherent vision should be built. This will require a deep investigation of the field, with the purpose of identifying the key problems, establishing a theoretical framework to understand what problems can be efficiently solved, and finally proposing a collection of decentralized algorithms.

3.1 Problem Identification

While selecting problems, two possible approaches may be taken: parsing the network analysis literature looking for properties that can be computed in a decentralized way, or wearing an "application hat": solving only problems coming from real, large-scale and decentralized scenarios. Following only the former, one risks to build a wonderful but otherwise useless theoretical cathedral, lacking any foundation on the practice; following only the latter, one risks to miss the general picture and propose just a bunch of algorithms.

3.2 Decentralization Issues

Going from a centralized approach to a decentralized one opens several exciting possibilities, but also introduces novel issues that are specific to distributed systems:

- *Off-line vs on-line*: Network analysis may be performed off-line (on static data) or on-line (on live networks). The latter case open the possibility of either adapting the protocols executed on the network, or even modifying the network itself in response to the results of the analysis. Careful attention to possible feedback loops between analysis and adaptation will be required.

- *Fault-tolerance*: Decentralization opens the possibility of terminating the evaluation of a network property even in the presence of failures. But this will not come for free: protocols will need to be appropriately designed to tolerate misbehavior. Correcting actions will be required, such as mechanisms for data replication or the exploitation of alternative paths to deliver messages.

- *Approximation vs exact computation* Linked to the previous point, but also related to the issue of scalability, a fundamental question is whether an approximate evaluation of a given property can be sufficient for a particular application.

- *Dynamic properties*: Some of the properties defined in the literature may suffer when networks are dynamic;

for example, large perturbations may be observed in betweenness centrality with the addition/removal of nodes. A thorough evaluation of the impact of dynamism on each particular property will be required.

Towards theoretical bounds Before starting reasoning about possible algorithms, an important question to be tackled is the following: "is it possible to formally define the class of properties that can be efficiently analyzed in a decentralized way, i.e. without concentrating the data in a single node"? The importance of this question must not be under-evaluated, as a clear answer will help to limit the development efforts only to problems that can effectively be solved.

Towards algorithms A broad portfolio of distributed algorithmic techniques can be applied to solve decentralize network analysis. We list here the most promising ones, with the obligatory disclaimer that this is only a limited list (probably biased by the author's background).

- *Peer-to-peer*: Recent research on peer-to-peer systems has generated several interesting protocols for the structural organization of (potentially large) overlay networks. We plan to leverage such results and build, whenever needed, appropriate structures aimed at facilitating the computation of network properties [6, 10]. It is important to note the distinction between the network to be analyzed and the overlay network that will be built to achieve this goal.

- *Gossip protocols*: In recent years the label "gossip" has been applied to an increasingly larger class of algorithms, going outside the original and limited field of information dissemination [4]. Gossip-based approaches exist now for information aggregation [8], overlay network management [6, 10] and clock synchronization [1]. Their distinctive features include relying on local information, being round-based and relatively simple, and having a bounded information transmission and processing complexity in each round. For these reasons, we believe that the gossip paradigm could be significantly applied to the field of decentralized network analysis.

- *Random walks*: While this technique can be seen as a special case of gossip-based protocols, it is worth mentioning on its own because of pre-existent works on this subject regarding the evaluation of network properties [12]. For example, betweenness centrality indeces may be evaluated using through Monte-Carlo methods between selected pairs of nodes.

4 Related work

A growing literature about the parallel evaluation of network properties exist, for both multi-processor and multicore systems. The lion's share of such literature is given by PageRank implementations [9], with even a large number of distributed versions [7, 14, 13].

Even centrality indexes such as betweennes centrality and closeness centrality have been the focus of parallel implementation; see [11, 3, 5, 2] for examples. In most of these cases, existing centralized algorithms are just moved to parallel systems; in some cases, approximate versions of such algorithms are discussed, based on Monte-Carlo simulation. Such approaches could be probably extended to distributed implementation.

5 Conclusions

The P2P paradigm has started a philosophical revolution on the Internet; it has become clear that collaboration between millions of users is possible. No matter if the first (and most successful) P2P application is the illegal sharing of copyrighted files: it is now common perception of designers and developers that several kinds of services may be successfully implemented by moving control from the center to the edges.

In this sense, the proposal of this paper is just another brick on the P2P wall: data representing large-scale networks are often distributed, so it is completely natural trying to understand whether it is possible to analyze them in a decentralized way.

The first and more important problem will be to clearly (and if possible, formally) identify the border between the set of problems that can be solved in a decentralized way, and the problems for which the solution will be not possible, even in an approximate form, because of the need of a global view.

Acknowledgments

Work supported by the project CASCADAS (IST-027807) funded by the FET Program of the European Commission.

References

[1] O. Babaoglu, T. Binci, M. Jelasity, and A. Montresor. Firefly-inspired heartbeat synchronization in overlay networks. In *Proceedings of the First IEEE International Conference on Self-Adaptive and Self-Organizing Systems (SASO 2007)*, Boston, USA, July 2007.

[2] D. A. Bader and K. Madduri. Parallel algorithms for evaluating centrality indices in real-world networks. In *Proc. The 35th International Conference on Parallel Processing (ICPP)*, pages 539–550, Columbus, OH, USA, Aug. 2006. IEEE Computer Society.

[3] U. Brandes and C. Pich. Centrality estimation in large networks. *International Journal of Bifurcation and Chaos*, 17(7):2303–2318, 2007.

[4] A. Demers, D. Greene, C. Hauser, W. Irish, J. Larson, S. Shenker, H. Sturgis, D. Swinehart, and D. Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the 6th Annual ACM Symposium on Principles of Distributed Computing (PODC'87)*, pages 1–12, Vancouver, British Columbia, Canada, August 1987. ACM Press.

[5] R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. In *Proc. of the 10th SIAM Workshop on Algorithm Engineering and Experiments (ALENEX08)*, 2008.

[6] M. Jelasity and O. Babaoglu. T-Man: Gossip-based overlay topology management. In S. A. Brueckner, G. Di Marzo Serugendo, D. Hales, and F. Zambonelli, editors, *Engineering Self-Organising Systems: Third International Workshop (ESOA 2005), Revised Selected Papers*, volume 3910 of *Lecture Notes in Computer Science*, pages 1–15. Springer-Verlag, 2006.

[7] M. Jelasity, G. Canright, and K. Engø-Monsen. Asynchronous distributed power iteration with gossip-based normalization. In A.-M. Kermarrec, L. Bougé, and T. Priol, editors, *Euro-Par*, volume 4641 of *Lecture Notes in Computer Science*, pages 514–525. Springer, 2007.

[8] M. Jelasity, A. Montresor, and O. Babaoglu. Gossip-based aggregation in large dynamic networks. *ACM Trans. Comput. Syst.*, 23(1):219–252, 2005.

[9] C. Kohlschütter, P.-A. Chirita, and W. Nejdl. Efficient parallel computation of pagerank. In M. Lalmas, A. MacFarlane, S. M. Rüger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors, *ECIR*, volume 3936 of *Lecture Notes in Computer Science*, pages 241–252. Springer, 2006.

[10] A. Montresor, M. Jelasity, and O. Babaoglu. Chord on demand. In *Proceedings of the 5th International Conference on Peer-to-Peer Computing (P2P 2005)*, pages 87–94, Konstanz, Germany, Aug. 2005. IEEE.

[11] S. Shi, J. Yu, G. Yang, and D. Wang. Distributed page ranking in structured p2p networks. In *ICPP*, pages 179–186. IEEE Computer Society, 2003.

[12] F. Spitzer. *Principles of Random Walk*. Springer, 1976.

[13] Q. Yang and S. Lonardi. A parallel edge-betweenness clustering tool for protein-protein interaction networks. *IJDMB*, 1(3):241–247, 2007.

[14] Y. Zhu, S. Ye, and X. Li. Distributed pagerank computation based on iterative aggregation-disaggregation methods. In O. Herzog, H.-J. Schek, N. Fuhr, A. Chowdhury, and W. Teiken, editors, *CIKM*, pages 578–585. ACM, 2005.