Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms

Proposal for a half-day tutorial at WWW'16

Francesco Bonchi* Gianmarco De Francisci Morales† Matteo Riondato‡

Topic and description

Identifying "important" nodes or edges in a graph is a fundamental task in network analysis, with many applications. Several measures, known as *centrality indices*, have been proposed over the years, each formalizing the concept of importance in a different way [15]. Centrality measures rely on graph properties to quantify importance. For example, betweenness centrality considers the fraction of shortest paths going through a node or edge, while the closeness centrality of a node is the average sum of the inverse of the distance to every other node. Other centrality measures use eigenvectors, random walks, degrees, or more complex properties, and can usually be extended to sets of nodes.

With the proliferation of huge networks with millions of nodes and billions of edges, the importance of having scalable algorithms for computing centrality indices has become more and more evident, and a number of contributions have been recently made, ranging from heuristics that perform extremely well in practice, to approximation algorithms with strong probabilistic guarantees, to scalable algorithms for the MapReduce platform. Moreover, the dynamic nature of many networks, i.e., the addition and removal of nodes and edges over time, dictates the need to keep the computed values of centrality up-to-date as the graph changes. These challenging problems have enjoyed enormous interest from the research community, resulting in several relevant approaches proposed recently to tackle them.

Our tutorial presents, in a unified framework, some of the different measures of centrality, and discusses the algorithms to compute them, both in an exact and in an approximate way, both in-memory and in a distributed fashion for the MapReduce framework of computation. Our tutorial represents an effort to ease the comparison between different measures, the different quality guarantees offered by approximation algorithms, and the different trade-offs and scalability behaviors characterizing distributed algorithms. We believe this unity of presentation is beneficial both for newcomers and for experienced researchers in the field.

*ISI Foundation, Turin Italy. `francescobonchi@acm.org`

†Aalto University, Helsinki, Finland. `gdfm@acm.org`

‡Two Sigma Investments, New York, NY, USA. `matteo@twosigma.com`. Main contact

Outline

The tutorial is structured in three main technical parts, plus a concluding part where we discuss future research directions. All the three technical parts will contains both theory and experimental results.

1. **Introduction: definitions and exact algorithms**

    1.1 The axioms of centrality [3]

    1.2 Definitions of centrality [15], including, but not limited to: betweenness, closeness, degree, eigenvector, harmonic, Katz, absorbing random-walk [13], and spanning-edge centrality [12].

    1.3 Betweenness centrality: exact algorithm [4] and heuristically-faster exact algorithms for betweenness centrality [7, 20].

    1.4 Exact algorithms for betweenness centrality in a dynamic graph [11, 14, 16].

    1.5 Exact algorithms for closeness centrality in a dynamic graph [19].

2. **Approximation algorithms**

    2.1 Sampling-based algorithm for closeness centrality [6].

    2.2 Betweenness centrality: almost-linear-time approximation algorithm [21], basic sampling-based algorithm [5], refined estimators [8], VC-dimension bounds for betweenness centrality [17].

    2.3 Approximation algorithms for betweenness centrality in dynamic graphs [1, 2, 9].

3. **Highly-scalable algorithms**

    3.1 GPU-based algorithms [18].

    3.2 Exact parallel streaming algorithm for betweenness centrality in a dynamic graph [10].

4. **Challenges and directions for future research**

Intended Audience

The tutorial is aimed at researchers interested in the theory and the applications of algorithms for graph mining and social network analysis.

We do not require any specific existing knowledge. The tutorial is designed for an audience of computer scientists who have a general idea of the problems and challenges in graph analysis. We will present the material in such a way that any advanced undergraduate student would be able to productively follow our tutorial. The tutorial starts from the basic definitions and progressively moves to more advanced algorithms, including sampling-based approximation algorithms and MapReduce algorithms, so that it will be of interest both to researchers new to the field and to a more experienced audience.

Duration: Half-day

Previous editions of the tutorial

The tutorial was not previously offered. We did not find any tutorial covering similar topics in the programs of recent relevant conferences.

Organizers

This tutorial is developed by Francesco Bonchi, Gianmarco De Francisci Morales, and Matteo Riondato. All three instructors will attend the conference.

Francesco Bonchi is Research Leader at the ISI Foundation, Turin, Italy, where he leads the "Algorithmic Data Analytics" group. He is also Scientific Director for Data Mining at Eurecat (Technological Center of Catalunya), Barcelona. Before he was Director of Research at Yahoo Labs in Barcelona, Spain, leading the Web Mining Research group.

His recent research interests include mining query-logs, social networks, and social media, as well as the privacy issues related to mining these kinds of sensible data.

He will be PC Chair of the 16th IEEE International Conference on Data Mining (ICDM 2016) to be held in Barcelona in December 2016. He is member of the ECML PKDD Steering Committee, Associate Editor of the newly created IEEE Transactions on Big Data (TBD), of the IEEE Transactions on Knowledge and Data Engineering (TKDE), the ACM Transactions on Intelligent Systems and Technology (TIST), Knowledge and Information Systems (KAIS), and member of the Editorial Board of Data Mining and Knowledge Discovery (DMKD). He presented a tutorial at ACM KDD'14.

Gianmarco De Francisci Morales is a Visiting Scientist at Aalto University. Previously he worked as a Research Scientist at Yahoo Labs Barcelona, and as a Research Associate at ISTI-CNR in Pisa. His research focuses on scalable data mining, with an emphasis on Web mining and data-intensive scalable computing systems. He is one of the lead developers of Apache SAMOA, an open-source platform for mining big data streams. He presented a tutorial on stream mining at IEEE BigData'14.

Matteo Riondato is a Research Scientist in the Labs group at Two Sigma Investments. Previously he was a postdoc at Stanford and at Brown. His dissertation on sampling-based randomized algorithms for data and graph mining received the Best Student Poster Award at SIAM SDM'14. His research focuses on exploiting advanced theory to develop practical algorithms for time series analysis, pattern mining, and social network analysis. He presented tutorials at ACM KDD'15, ECML PKDD'15, and ACM CIKM'15.

Support materials

We are developing a mini-website at `http://matteo.rionda.to/centrtutorial/`. It will contain the abstract of the tutorial, a detailed outline with short a description of each item of the outline, a full list of references with links to electronic editions, a list of software packages implementing the algorithms, and the slides used in the tutorial presentation. A preliminary version of the website will be available 15 days after the tutorial is accepted. A preliminary version of the slides will be available 30 days before the conference, or in any case by any deadline given to us by the conference organizers, and the final version will be available 15 days before the conference.

References

[1] E. Bergamini and H. Meyerhenke. Fully-dynamic approximation of betweenness centrality. ESA'15, 2015.

[2] E. Bergamini, H. Meyerhenke, and C. L. Staudt. Approximating betweenness centrality in large evolving networks. ALENEX'15, 2015.

[3] P. Boldi and S. Vigna. Axioms for centrality. *Internet Mathematics*, 10(3–4):222–262, 2014.

[4] U. Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2): 163–177, 2001.

[5] U. Brandes and C. Pich. Centrality estimation in large networks. *Int. J. Bifurcation and Chaos*, 17(7):2303–2318, 2007.

[6] D. Eppstein and J. Wang. Fast approximation of centrality. *J. Graph Algorithms Appl.*, 8(1):39–45, 2004.

[7] D. Erdős, V. Ishakian, A. Bestavros, and E. Terzi. A divide-and-conquer algorithm for betweenness centrality. SDM'15, 2015.

[8] R. Geisberger, P. Sanders, and D. Schultes. Better approximation of betweenness centrality. ALENEX'08, 2008

[9] M. Kas, M. Wachs, K. M. Carley, and L. R. Carley. Incremental algorithm for updating betweenness centrality in dynamically growing networks. ASONAM'13, 2013.

[10] N. Kourtellis, G. D. F. Morales, and F. Bonchi. Scalable online betweenness centrality in evolving graphs. *IEEE Trans. Knowl. Data Eng.*, 27(9):2494–2506, 2015.

[11] M.-J. Lee, J. Lee, J. Y. Park, R. H. Choi, and C.-W. Chung. QUBE: A quick algorithm for updating betweenness centrality. WWW'12, 2012.

[12] C. Mavroforakis, R. Garcia-Lebron, I. Koutis, and E. Terzi. Spanning edge centrality: Large-scale computation and applications. WWW'15, 2015.

[13] C. Mavroforakis, M. Mathioudakis, and A. Gionis. Absorbing random-walk centrality: Theory and algorithms. CoRR/abs 1509.02533, 2015.

[14] M. Nasre, M. Pontecorvi, and V. Ramachandran. Betweenness centrality–incremental and faster. MFCS'14, 2014.

[15] M. E. J. Newman. *Networks – An Introduction.* Oxford University Press, 2010.

[16] M. Pontecorvi and V. Ramachandran. A faster algorithm for fully dynamic betweenness centrality. CoRR/abs 1506.05783, 2015.

[17] M. Riondato and E. M. Kornaropoulos. Fast approximation of betweenness centrality through sampling. *Data Mining Knowl. Disc.*, (to appear), 2015.

[18] A. E. Sariyüce, K. Kaya, E. Saule, and Ü. V. Çatalyürek. Betweenness centrality on gpus and heterogeneous architectures. In *Proceedings of the 6th Workshop on General Purpose Processor Using Graphics Processing Units*, pages 76–85. ACM, 2013.

[19] A. E. Sariyuce, K. Kaya, E. Saule, and U. V. Catalyurek. Incremental algorithms for closeness centrality. BigData'13, 2013.

[20] A. E. Sarıyüce, E. Saule, K. Kaya, and U. V. Çatalyürek. Shattering and compressing networks for betweenness centrality. SDM'13, 2013.

[21] Y. Yoshida. Almost linear-time algorithms for adaptive betweenness centrality using hypergraph sketches. KDD'14, 2014.