Large-scale Network Analytics: Diffusion-based Computation of Distances and Geometric Centralities

Paolo Boldi
Dipartimento di Informatica Università degli Studi di Milano — Italy

ABSTRACT

Given a large complex network, which of its nodes are more central? This question emerged in many contexts (e.g., sociology, psychology and computer science), and gave rise to a large range of proposed centrality measures. Providing a sufficiently general and mathematically sound classification of these measures is challenging: on one hand, it requires that one can suggest some simple, basic properties that a centrality measure should exhibit; on the other hand, it calls for innovative algorithms that allow an efficient computation of these measures on large real networks. HyperBall is a recently proposed tool that accesses the graph in a semi-streaming fashion and is at the same time able to compute the distance distribution and to approximate all geometric (i.e., distance-based) centralities. It uses a very small amount of core memory, thanks to the application of HyperLogLog counters, and exhibits high, guaranteed accuracy.

Given an unweighted directed graph $G = (V, E)$, let us write $d(x, y)$ for the length of the shortest path from $x \in V$ to $y \in V$ (or ∞, if there is no such path) and let $B_t(y) = \{x \in V \mid d(x, y) \leq t\}$. A *geometric centrality* is a centrality measure c that is expressible as a function of the sequence $\langle |B_t(y)| \rangle_{t \in \mathbf{N}}$. More precisely, there must be two functions $h : \mathbf{N} \times \mathbf{N} \times \mathbf{R} \to \mathbf{R}$ and $g : \mathbf{R} \to \mathbf{R}$ such that $c(x) = \lim_{t \to \infty} g(H_t(x))$ where $H_0(x) = h(0, 0, 1)$ and $H_{t+1}(x) = h(t + 1, |B_{t+1}(x)|, H_t(x))$.

For example, letting $h(0, -, -) = 0$, $h(t + 1, x, v) = v + (t + 1)x$ and $g(v) = 1/v$ gives the standard closeness centrality as defined in [2], whereas harmonic centrality [5] is obtained by $h(0, -, -) = 0$, $h(t + 1, x, v) = v + x/(t + 1)$ and $g(v) = v$.

It is interesting to observe that $B_t(y)$ can be computed iteratively by accessing the graph in a semi-streaming fashion (i.e., only scanning sequentially the list of arcs at every iteration); in fact, $B_0(y) = \{y\}$ and $B_{t+1}(y) = \{y\} \cup_{(x,y) \in E} B_t(x)$. Moreover, the number of iterations needed to reach stability is the length of the longest shortest path in G (the graph diameter, if G is strongly connected). The advantage of this approach (with respect to an all-pairs shortest-path computation, or a sampling technique based on a few breadth-first traversals) is that it behaves much better even on highly disconnected graphs and it does not require random access to the graph, so it is more cache- and compression-friendly.

HyperBall [6] leverages the newly discovered algorithms based on HyperLogLog counters [3] to store an approximated version of $B_t(y)$ for every node y, making it possible to approximate geometric centralities at a very high speed and with high accuracy. While the application of similar algorithms for the approximation of closeness was attempted in the MapReduce framework, our exploitation of HyperLogLog counters reduces exponentially the memory footprint, and it is about two orders of magnitude faster, paving the way for in-core processing of networks with a hundred billion nodes using "just" 2 TiB of RAM. Moreover, HyperBall is inherently parallelizable, and scales linearly with the number of available cores.

In this talk, I will first offer a bird's-eye view of centralities, trying to provide a principled taxonomy based on an axiomatic approach; then I will discuss the computational issue involved and describe the diffusive solution adopted by HyperBall; finally, as a case study, I will report the results of the execution of HyperBall for the computation of the distance distribution on some snapshots of the Facebook network, that allowed us to obtain the by-now well-known "Four Degrees of Separation" result [1, 4].

WWW'15 Companion May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2741703.

1. REFERENCES

[1] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *ACM Web Science 2012: Conf. Proc.*, pages 45–54. ACM Press, 2012. Best paper award.

[2] Alex Bavelas. Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.*, 22(6):725–730, 1950.

[3] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proc. of the 20th International Conf. on World Wide Web*, pages 625–634. ACM, 2011.

[4] Paolo Boldi and Sebastiano Vigna. Four degrees of separation, really. In *Proc. of the 2012 International Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1222–1227. IEEE Computer Society, 2012.

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

[6] Paolo Boldi and Sebastiano Vigna. In-core computation of geometric centralities with HyperBall: A hundred billion nodes and beyond. In *Proc. of 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW 2013)*. IEEE, 2013.