Personalized PageRank for Local Community Detection

Fan Chen (joint work with Yini Zhang and Karl Rohe)

SIAM Workshop on Network Science (NS18)

UNIVERSITY OF WISCONSIN-MADISON

Introduction

Much of the literature on graph sampling has treated the entire social graph, or all of the people in it, as the target population. However, in many settings, the target population is only a particular community in the massive graph. This corresponds to identifying potentially small communities in a massive network.

In such an application, the graph is useful for two primary reasons. First, via link tracing, we can find potential members of the target population. Second, the graph connections are informative for identifying community membership.

The Algorithm

Input: adjacency matrix A, seed node v_0, and the desired size of local cluster n.

1. Calculate the approximate PPR vector p [Andersen et al, 2006].
2. Adjust the PPR vector p by node degrees, $p_v^* \leftarrow p_v/d_v$.
3. Rank all vertices according to the adjusted PPR vector p^*.

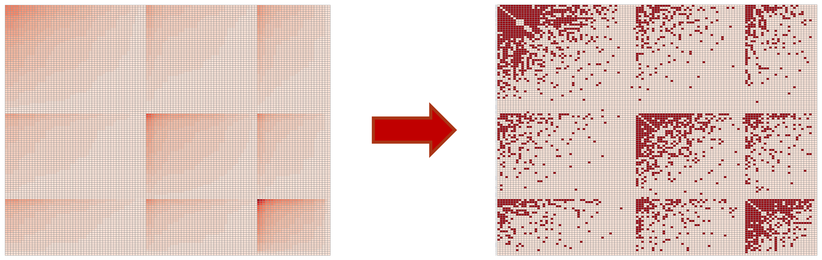
Output: local cluster -- n top-ranking nodes.

Methodology

Statistical Model

To understand the effects of heterogeneous node degrees, we use a statistical model with underlying community structure, the Degree-Corrected Stochastic Blockmodel (DC-SBM) [Karren and Newman, 2011]. Under the DC-SBM with K underlying blocks and N nodes, each vertex belongs to one block (akin to planted solution). If two vertices u, v belong to block i, j, then the DC-SBM assumes the presence of an edge between u and v depends only on an block connectivity matrix $B \in \mathbb{R}^{K\times K}$ and their degree parameters θ_u, θ_v,

$$\mathbb{P}(u \leftrightarrow v) = \theta_u \theta_v B_{ij}.$$



Element-wise Eigenvector Perturbation Bounds

Recent advances in eigenvector perturbation bound enable an element-wise control of the leading eigenvectors of a random Markovian transition matrix, provided sufficiently large size and dense enough support.

Theoretical Results

Characterizations of PPR

Theorem 1. Under the population DC-SBM, the PPR of each vertex is the product of its degree parameter and the PPR for the block it belongs to.

Corollary 2. Adjusting PPR by node degrees yields an unbiased estimator of local community, which ranks vertices from the same block as the seed on the top.

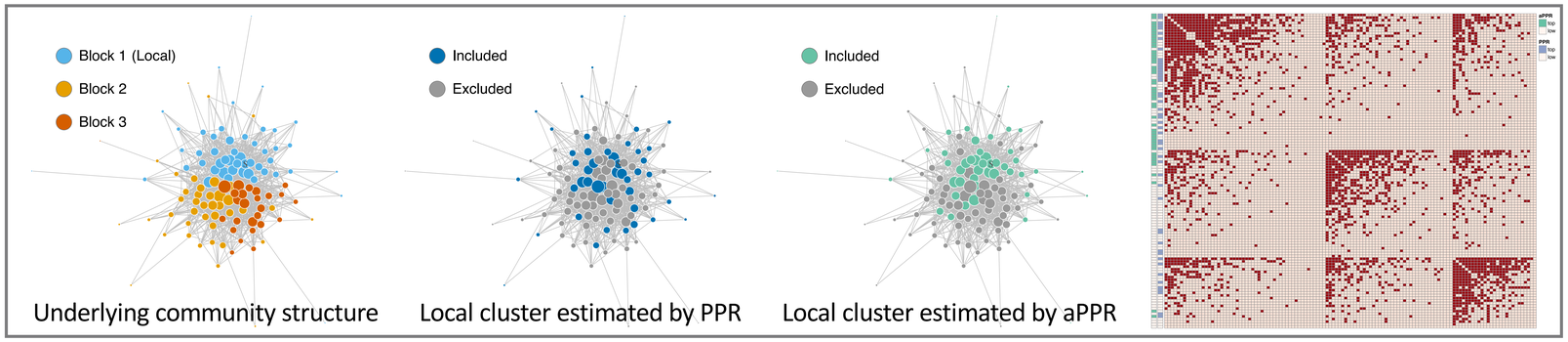
Theorem 3. If the graph is generated by DC-SBM, and the graph is large and dense enough, $\bar{d} \gtrsim \mathcal{O}(\log N)$, then the PPR vector is entry-wise close to its population (expectation) with high probability.

An Application to Twitter

Local clusters with Nate Silver (@NateSilver538) as the seed

By PPR	By rPPR	Description
Donald J. Trump	Zach Thrun	Creator of Aircamp Games.
FiveThirtyEight	Gina	still figuring this tweet thing out
The New York Times	Renard Sexton	Emory Asst Prof // Contributor at FiveTh...
President Trump	Chris A	just trying to be a good husband ...
Pew Research Center	Brett Marty	Director, sometimes photographer
The Onion	Brian D. Silver	Michigan State University, Emeritus prof ...
Ezra Klein	GOP Delegate Math	Corrections and clarifications.
Nate Cohn	Kat Reid	Project managing ... Previous @Yahoo ...
Ariel Edwards-Levy	F&M Opinion Research	F&M Center for Opinion Research ...
(((Harry Enten)))	PSRAI	Conducting public opinion research ...
David Leonhardt	Stephanie Roos	Roots in DC, life in NYC.
Hillary Clinton	Jerilyn Bowers	Fundraiser and PR pro for MDI Bio Lab ...

Numerical Example



Underlying community structure | Local cluster estimated by PPR | Local cluster estimated by aPPR

References

Andersen, R., Fan, C., & Lang, K. (2006). Local Graph Partitioning using PageRank Vectors. *IEEE Symposium on Foundations of Computer Science* (pp.475-486). IEEE Computer Society.

Karrer, B., & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, *83*(1), 016107.