Personalized PageRank and Local Community Detection

Fan Chen
(Joint work with Yini Zhang and Karl Rohe)
University of Wisconsin-Madison

2018 SIAM Annual Meeting, Portland

Outlines

- Background: local community detection
- Method: Personalized PageRank
- Findings:
  (1) what it is doing under a statistical model, and
  (2) a simple bias adjustment
  (3) confidence of this estimation

Background

- Massive network brings challenges to computation
  - Twitter users (336m active monthly)
  - Academic collaborations (17m faculty members and graduate students)
- Many times, target population is a small community
  - Political reporters
  - Linear algebra in network computations
- **Goal:** identify a small community efficiently in time/memory

Idea: use random walk from a seed

- Starting from **seed** node, walk to a neighbor uniformly at random
- Don't go **too** far!
- Teleportation probability α
- At each step,

$$\mathbb{P}(\text{return to seed node}) = \alpha$$
$$\mathbb{P}(\text{walk to a neighbour}) = 1 - \alpha$$

- Use the stationary distribution

Algorithm: Personalized PageRank (PPR)

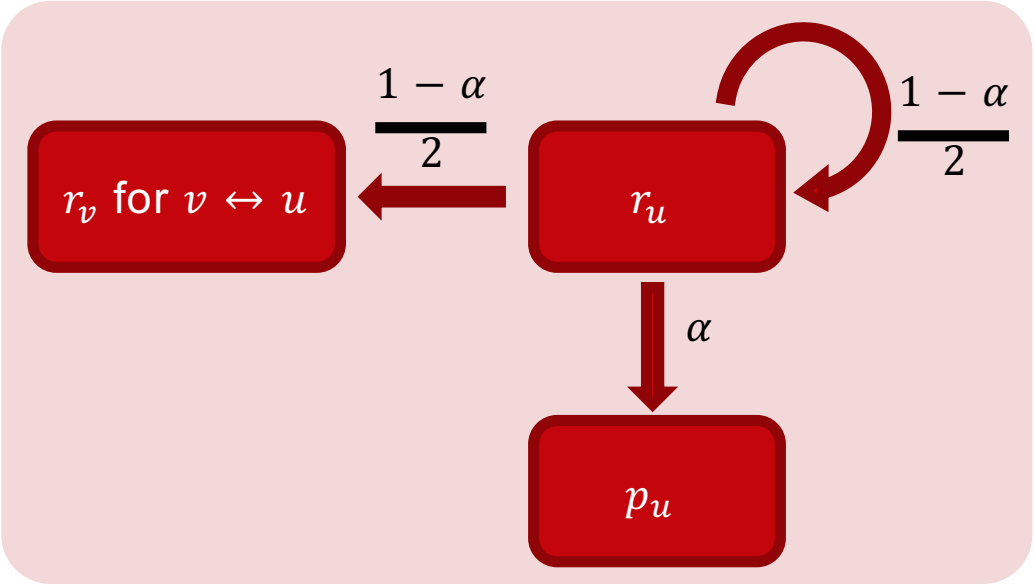
- Adjacency matrix $A \in \{0,1\}^{N \times N}$
- Graph transition P (i.e. A normalized by column sum)
- PPR vector is the leading eigenvector of

$$\alpha \Pi + (1 - \alpha) P$$

where Π has all 1 in the first row and 0 elsewhere

PPR can be quickly approximated

- Initialize a residual $r = (1,0,\dots,0)$, and $p = (0,\dots,0)$
- While there exists a node u with large enough residual r_u, distribute r_u in three ways $\alpha, \frac{1-\alpha}{2}, \frac{1-\alpha}{2}$ into [Andersen et al, 2006]
  - p_u
  - r_u
  - r_v, equally for $u \leftrightarrow v$



Is PPR good? Or Best?

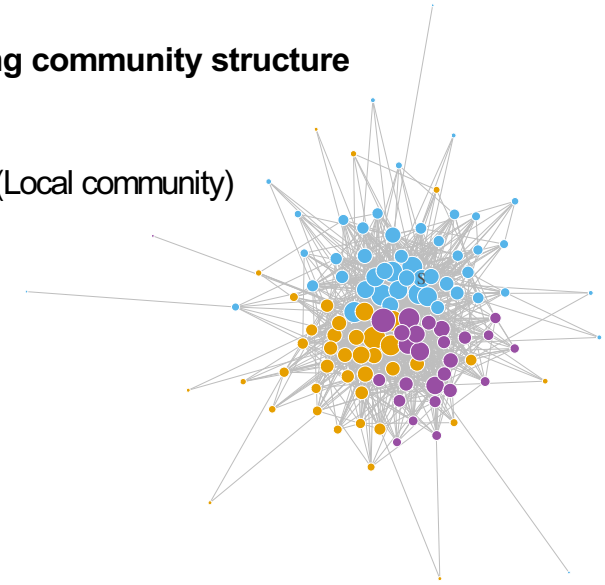
Nate Silver (@NateSilver538) April '18

User	Description	Followers
Donald J. Trump	45th President of the United States of America🇺🇸	51385809
FiveThirtyEight	The home of Nate Silver's FiveThirtyEight on Twitter.	957788
The New York Times	Where the conversation begins. Follow for breaking news…	41985496
President Trump	45th President of the United States of America…	22997330
Pew Research Center	Nonpartisan, non-advocacy data and analysis on the issues…	359427
The Onion	America's Finest News Source.	11407493
Ezra Klein	Founder and editor-at-large, https://t.co/5gESirESRH…	2498243
Nate Cohn	I write for The New York Times at @UpshotNYT…	178721
Ariel Edwards-Levy	Reporter and polling editor @HuffPostPol, covering …	32036
(((Harry Enten)))	Son of a man who was far from perfect, but I loved him…	114161
David Leonhardt	Op-Ed columnist, The New York Times …	112092
Hillary Clinton	2016 Democratic Nominee, SecState, Senator, hair icon…	22658733

Is PPR good? Or best?

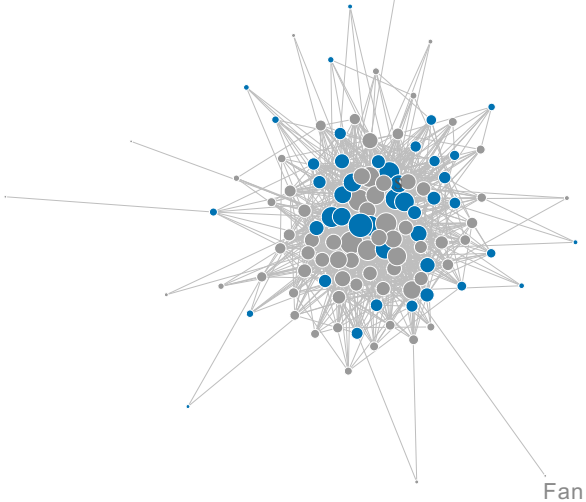
Underlying community structure

- Block 1 (Local community)
- Block 2
- Block 3



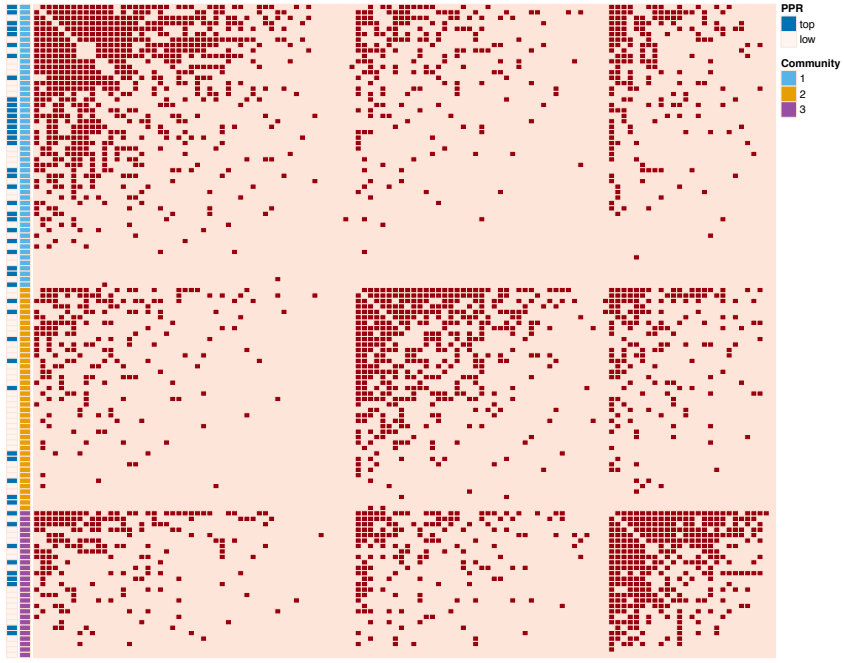
Local community estimated by PPR

- Included
- Excluded



10/41

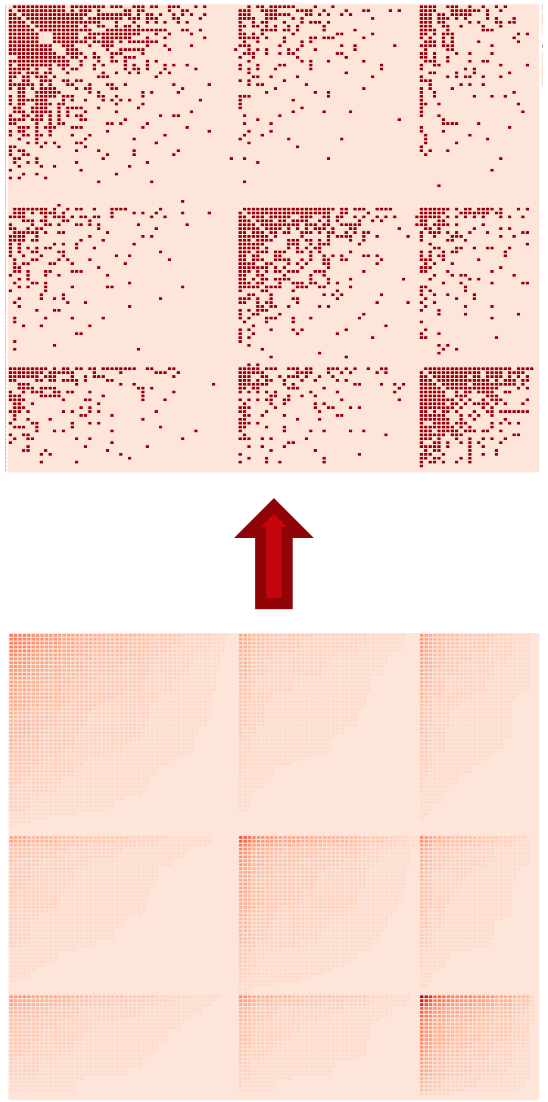
9/27



Use a statistical model: Blockmodel

- K underlying blocks and N nodes
- **Planted solution**: each vertex belongs to one block
- Block connectivity matrix $B \in \mathbb{R}^{K \times K}$
- Degree parameters θ_v
- If u, v belong to block i, j, the Degree-Corrected Stochastic Blockmodel (DC-SBM) says [Karren and Newman, 2011]

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



PPR is biased toward high degree nodes

- $\tilde{p}$ is block-wise PPR vector, that is the PPR vector corresponding to weighted adjacency matrix B
- Under population DC-SBM, the PPR of each vertex is the product of its degree parameter and the PPR for its block, $p_v = \theta_v \tilde{p}_i$.
- PPR is confounded by node degree

But, a simple adjustment works

- Adjust PPR by node degree,

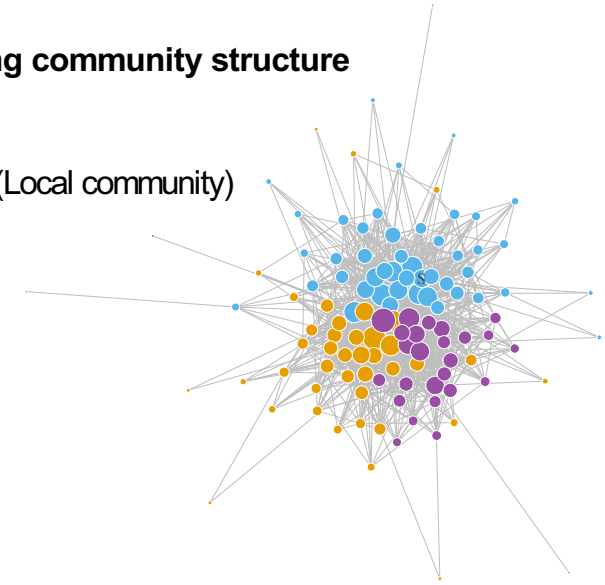
$$p_v^* = \frac{p_v}{d_v}$$

- Adjusted PPR (aPPR) guarantees to rank local block on top
- If the network is generated from DC-SBM, and if the graph is large and dense enough, $d \gtrsim \mathcal{O}(\log N)$, then the PPR vector is **entrywise** close to its population (expectation) with high probability.

Adjusted PPR finds local community

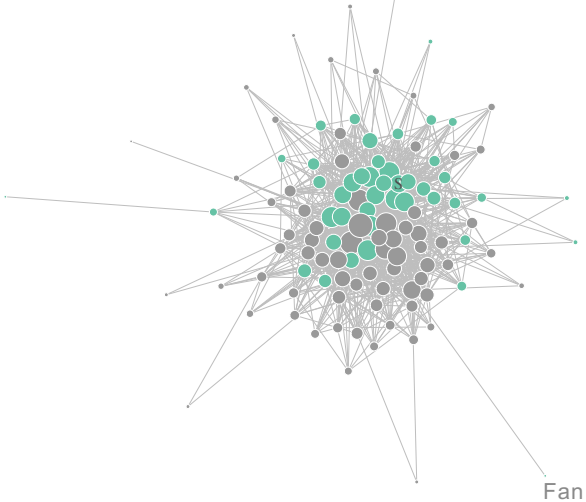
Underlying community structure

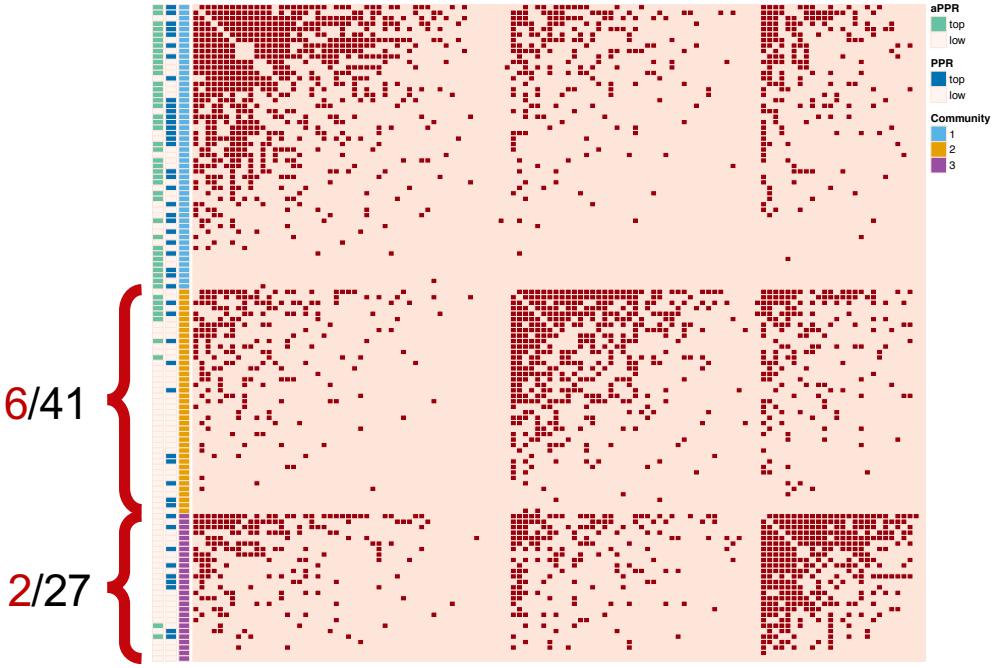
- Block 1 (Local community)
- Block 2
- Block 3



Local community estimated by aPPR

- Included
- Excluded





Example: simply adjusted PPR is noisy

Nate Silver (@NateSilver538) April '18

User	Rank by PPR	Rank By aPPR
Donald J. Trump	2	5490
FiveThirtyEight	3	2425
The New York Times	4	5482
President Trump	5	5175
Pew Research Center	6	1623
The Onion	7	4720
Ezra Klein	8	3538
Nate Cohn	9	1191
Ariel Edwards-Levy	10	454
(((Harry Enten)))	11	951
David Leonhardt	12	949
Hillary Clinton	13	5241

Solution: regularization

- Node degrees are **noisy** empirically
- A regularized adjustment:

$$p_v^* = \frac{p_v}{d_v + \tau}$$

- Regularizer τ is set to average node degree [Tai and Rohe, 2011]

Example: regularized PPR is localized

Nate Silver (@NateSilver538)

April '18

User	Description	Rank	Followers
Renard Sexton	Princeton Postdoc // Emory Asst Prof // Contributor at FiveThirty…	3	162
Brett Marty	Director, sometimes photographer @specfilms and @youthfilm2016	5	157
Brian D. Silver	Michigan State University, Emeritus Prof. …	6	190
GOP Delegate Math	Corrections and clarifications re: GOP delegate allocation rules.	7	142
Kat Reid	Project managing all the things @Splunk. Previous @Yahoo. …	12	226

Thanks!

- Q&A

- **Reference**
  F. Chen, Y. Zhang, K. Rohe. Personalized PageRank for simultaneous sampling and estimation of local clusters in massive Blockmodel graphs.