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 $\alpha$
- 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 $\Pi$ has all 1 in the first row and 0 elsewhere
PPR can be quickly approximated

- Initialize a residual $r = (1, 0, \ldots, 0)$, and $p = (0, \ldots, 0)$
- 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

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

Local community estimated by PPR

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 $\theta_v$
- 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

- $\tilde{\theta}$ 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,
  \[ \rho_v = \theta_v \tilde{\rho}_i. \]

- 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

6/41

2/27
Example: simply adjusted PPR is noisy

Nate Silver (@NateSilver538)

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

- Node degrees are **noisy** empirically
- A regularized adjustment:
  \[ p_v^* = \frac{p_v}{d_v + \tau} \]
- Regularizer \( \tau \) is set to average node degree [Tai and Rohe, 2011]
Example: regularized PPR is localized

Nate Silver (@NateSilver538)  

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

• Q&A

• Reference