Personalized PageRank based Community Detection

Joint work with C. Seshadhri, Joyce Jiyoung Whang, and Inderjit S. Dhillon, supported by NSF CAREER 1149756-CCF

David F. Gleich
Purdue University

Code bit.ly/dgleich-codes
Today’s talk

1. Personalized PageRank based community detection
2. Conductance, Egonets, and Network Community Profiles
3. Egonet seeding
4. Improved seeding
A community is a set of vertices that is denser inside than out.
250 node GEOP network in 2 dimensions
250 node GEOP network in 2 dimensions
We can find communities using *Personalized PageRank (PPR)* [Andersen et al. 2006]

PPR is a Markov chain on nodes

1. with probability $\alpha$, follow a random edge
2. with probability $1-\alpha$, restart at a seed

aka *random surfer*
aka *random walk with restart*
unique stationary distribution
Personalized PageRank community detection

1. Given a seed, approximate the stationary distribution.
2. Extract the community.

Both are local operations.
Demo!
Conductance communities

*Conductance* is one of the most important community scores \[\text{[Schaeffer07]}\]

The conductance of a set of vertices is the ratio of edges leaving to total edges:

\[
\phi(S) = \frac{\text{cut}(S)}{\min(\text{vol}(S), \text{vol}(\bar{S}))}
\]

Equivalentely, it’s the probability that a random edge leaves the set.

Small conductance $\iff$ Good community

\[
\begin{align*}
\text{cut}(S) &= 7 \\
\text{vol}(S) &= 33 \\
\text{vol}(\bar{S}) &= 11 \\
\phi(S) &= 7/11
\end{align*}
\]
Informally

Suppose the seeds are in a set of good conductance, then the personalized PageRank method will find a set with conductance that’s nearly as good.

... also, it’s really fast.
# G is graph as dictionary-of-sets
alpha=0.99
tol=1e-4

x = {} # Store x, r as dictionaries
r = {} # initialize residual
Q = collections.deque() # initialize queue
for s in seed:
    r(s) = 1/len(seed)
    Q.append(s)
while len(Q) > 0:
    v = Q.popleft() # v has r[v] > tol*deg(v)
    if v not in x: x[v] = 0.
    x[v] += (1-alpha)*r[v]
    mass = alpha*r[v]/(2*len(G[v]))
    for u in G[v]: # for neighbors of u
        if u not in r: r[u] = 0.
        if r[u] < len(G[u])*tol and \
            r[u] + mass >= len(G[u])*tol:
            Q.append(u) # add u to queue if large
        r[u] = r[u] + mass
    r[v] = mass*len(G[v])
Demo 2!
Problem 1, which seeds?
Problem 2, not fast enough.
Neighborhoods are good communities
Verext

Neighborhoods are good communities

Egonents?

... in graphs that look like social and information networks
Vertex neighborhoods or Egonets

The induced subgraph of set a vertex its neighbors

Prior research on egonets of social networks from the “structural holes” perspective [Burt95, Kleinberg08]. Used for anomaly detection [Akoglu10], community seeds [Huang11, Schaeffer11], overlapping communities [Schaeffer07, Rees10].
Simple version of theorem

If global clustering coefficient = 1, then the graph is a disjoint union of cliques.

Vertex neighborhoods are optimal communities!
Theorem

**Condition** Let graph $G$ have clustering coefficient $\kappa$ and have vertex degrees bounded by a power-law function with exponent $\gamma$ less than 3.

**Theorem** Then there exists a vertex neighborhood with conductance

\[ \leq \frac{4(1 - \kappa)}{(3 - 2\kappa)} \]
Confession
The theory is weak

\[ \phi(S) \leq 4(1 - \kappa)/(3 - 2\kappa) \]

This bound is useless unless \( \kappa \geq 1/2 \)

<table>
<thead>
<tr>
<th>Graph</th>
<th>Verts</th>
<th>Edges</th>
<th>( \kappa )</th>
<th>( \tilde{C} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ca-AstroPh</td>
<td>17903</td>
<td>196972</td>
<td>0.318</td>
<td>0.633</td>
</tr>
<tr>
<td>email-Enron</td>
<td>33696</td>
<td>180811</td>
<td>0.085</td>
<td>0.509</td>
</tr>
<tr>
<td>cond-mat-2005</td>
<td>36458</td>
<td>171735</td>
<td>0.243</td>
<td>0.657</td>
</tr>
<tr>
<td>arxiv</td>
<td>86376</td>
<td>517563</td>
<td>0.560</td>
<td>0.678</td>
</tr>
<tr>
<td>dblp</td>
<td>226413</td>
<td>716460</td>
<td>0.383</td>
<td>0.635</td>
</tr>
<tr>
<td>hollywood-2009</td>
<td>1069126</td>
<td>56306653</td>
<td>0.310</td>
<td>0.766</td>
</tr>
<tr>
<td>fb-Penn94</td>
<td>41536</td>
<td>1362220</td>
<td>0.098</td>
<td>0.212</td>
</tr>
<tr>
<td>fb-A-oneyear</td>
<td>1138557</td>
<td>4404989</td>
<td>0.038</td>
<td>0.060</td>
</tr>
<tr>
<td>fb-A</td>
<td>3097165</td>
<td>23667394</td>
<td>0.048</td>
<td>0.097</td>
</tr>
<tr>
<td>soc-LiveJournal1</td>
<td>4843953</td>
<td>42845684</td>
<td>0.118</td>
<td>0.274</td>
</tr>
<tr>
<td>oregon2-010526</td>
<td>11461</td>
<td>32730</td>
<td>0.037</td>
<td>0.352</td>
</tr>
<tr>
<td>p2p-Gnutella25</td>
<td>22663</td>
<td>54693</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>as-22july06</td>
<td>22963</td>
<td>48436</td>
<td>0.011</td>
<td>0.230</td>
</tr>
<tr>
<td>itdk0304</td>
<td>190914</td>
<td>607610</td>
<td>0.061</td>
<td>0.158</td>
</tr>
</tbody>
</table>

Collaboration networks
\( \kappa \sim [0.1 - 0.5] \)

Social networks
\( \kappa \sim [0.05 - 0.1] \)

Tech. networks
\( \kappa \sim [0.005 - 0.05] \)
We view this theory as "intuition for the truth"
Empirical Evaluation using Network Community Profiles

Minimum conductance for any community of the given size

Approximate canonical shape found by Leskovec, Lang, Dasgupta, and Mahoney

Holds for a variety of approximations to conductance.
Minimum conductance for any community neighborhood of the given size.
Not just one graph

arXiv – 86k verts, 500k edges

soc-LiveJournal – 5M verts, 42M edges

15 more graphs available

www.cs.purdue.edu/~dgleich/codes/neighborhoods
Filling in the Network Community Profile

Minimum conductance for any community neighborhood of the given size

Facebook Sample - 1.1M verts, 4M edges

We are missing a region of the NCP when we just look at neighborhoods

David Gleich · Purdue

MLG2013
Filling in the Network Community Profile

Minimum conductance for any community of the given size

7807 seconds

This region fills when using the PPR method (like now!)
Am I a good seed?
Locally Minimal Communities

“My conductance is the best locally.”

\[ \phi(N(v)) \leq \phi(N(w)) \]
for all \( w \) adjacent to \( v \)

In Zachary’s Karate Club network, there are four locally minimal communities, the two leaders and two peripheral nodes.
Locally minimal communities capture extremal neighborhoods

Red dots are conductance and size of a locally minimal community.

Usually about 1% of # of vertices.

The red circles – the best local mins – find the extremes in the egonet profile.
Filling in the NCP
Growing locally minimal comm.

PPR growing only locally min communities, seeded from entire egonet

3 seconds
283 seconds
7807 seconds
But there’s a small problem. Most people want to cover a network with communities! We just looked at the best.
The coverage of egonet-grown communities is really bad.

Facebook network

With a conductance of 0.1 (not so good) we only cover 1% of the vertices in the network.
1. Extract part of the graph that might have overlapping communities.

2. Compute a partitioning of the network into many pieces (think $\sqrt{n}$) using Graclus.

3. Find the center of these partitions.

4. Use PPR to grow egonets of these centers.
A good partitioning helps

Flickr social network
2M vertices
22M edges

We can cover 95% of network with communities of cond. ~0.15.
And helps to find real-world overlapping communities too.

Using datasets from Yang and Leskovec (WDSM 2013) with known overlapping community structure.

Our method outperform current state of the art overlapping community detection methods. Even randomly seeded!

Our seed set expansion method was the only method that worked on all of the problems. Also, our method is faster than both Bigclam and Demon.

Our seed set expansion algorithm is also easy to parallelize because each seed can be expanded independently. This property indicates that the runtime of the seed set expansion method could be further reduced in a multi-threaded version. Also, we can use any other high quality partitioning scheme instead of Graclus including those with parallel and distributed implementations [25]. Perhaps surprisingly, the major difference in cost between using Graclus centers for the seeds and the other seed choices does not result from the expense of running Graclus. Rather, it arises because the personalized PageRank expansion technique takes longer for the seeds chosen by Graclus and spread hubs. When the PageRank expansion method has a larger input set, it tends to take longer, and the input sets we provide for the spread hubs and Graclus seeding strategies are the neighborhood sets of high degree vertices.

Another finding that emerges from our results is that using random seeds outperforms both Bigclam and Demon. We believe there are two reasons for this finding. First, random seeds are likely to be in some set of reasonable conductance as also discussed by Andersen and Lang [5]. Second, and importantly, a recent study by Abrahao [2] showed that personalized PageRank clusters are topologically similar to real-world clusters [2]. Any method that uses this technique will find clusters that look real.

Finally, we wish to address the relationship between our results and some prior observations on overlapping communities. The authors of Bigclam found that the dense regions of a graph reflect areas of overlap between overlapping communities. By using a conductance measure, we ought to find only these dense regions – however, our method produces much larger communities that cover the entire graph. The reason for this difference is that we use the entire vertex neighborhood as the restart for the personalized PageRank expansion routine. We avoid seeding exclusively inside a dense region by using an entire vertex neighborhood as a seed, which grows the set beyond the dense region. Thus, the communities we find likely capture a combination of communities given by the egonet of the original seed node. To expand on this point, in experiments we omit due to space, we found that seeding solely on the node itself – rather than using...
Conclusion & Discussion & References

PPR community detection is fast
[Andersen et al. FOCS06]

PPR communities look real
[Abrahao et al. KDD2012; Zhu et al. ICML2013]

Egonet analysis reveals basis of NCP

Partitioning for seeding yields high coverage & real communities.

“Caveman” communities?

Best conductance cut at intersection of communities?

Gleich & Seshadhri
KDD2012

Whang, Gleich & Dhillon
CIKM2013

PPR Sample
bit.ly/18khz05

Egonet seeding
bit.ly/dgleich-code

David Gleich · Purdue
MLG2013
Proof Sketch

1) Large clustering coefficient
   ⇒ many wedges are closed

2) Heavy tailed degree dist
   ⇒ a few vertices have a very large degree

3) Large degree ⇒ \( O(d^2) \) wedges ⇒ “most” of wedges

Thus, there must exist a vertex with a high edge density ⇒ “good” conductance

*Use the probabilistic method to formalize*