iTTop - Interaction Based Topic Centric Community Discovery on Twitter

@denzil_correa, @ashish_sureka and @mayank_pundir
Research Motivation

- Meteoric Rise of Social Media
- Formation of Communities
- Twitter - A popular microblog service
Real World Use Cases

- **Digital Marketing**
  - Identify target customer groups
  - Plan product releases & announcements

- **Law Enforcement Agencies**
  - Identify xenophobic communities
  - What do these communities talk about?
Technical Challenges

- Large Scale - *Finding a needle in a haystack*

- Heterogeneous Interactions

- Community Evaluation
Large Scale Networks
Local Modularity Algorithm

- Aaron Clauset 2008

- Does not load complete graph into memory

- Maximizes a structural based objective function

- Finds concentric communities around nodes
Local Modularity Algorithm

- **Steps**
  - Start with Seed Node; initialize community $C = \{\text{seed node}\}$
  - Expand Node to obtain candidate nodes
  - For each candidate node, calculate a value based on an objective function
  - Add the node which maximizes the objective function to community $C$
  - Repeat all steps till stopping condition reached
Best !!!

R = 0.4

R = 0.95

R = 0.6

R = 0.8

u1

u2

u4

u3

Seed

Community
Objective Function

- Local Modularity $R$ [Aaron Clauset 2008]

- $R$ is proportional to the ratio of “Number of edges within the community” to that of “Number of edges outside the community”

- Rewards nodes which build community structure
Advantages of Local Modularity

- Visits only those nodes which are potential candidates
- Saves memory and processing power
Heterogeneous Interactions
Follower network?

- Follower-Followee network is a bust! [Huberman et al. 2008]

- True social network lies beneath this network in terms of interactions
InfoChimps - StrongLinks

- REST API service

- Captures interaction between users
  - Frequency of Retweets
  - Frequency of @-mentions
  - Frequency of Reply or Reply to a Retweet

- Helps us know the strongest connection
Connecting the dots...
iTTop - Proposed Algorithm

- Warm Start
- Expand
- Filter
- Iterate
Input Topic Keyword T
Warm Start - Sow the seeds!

- Twitter Users have a *Bio*

- Search Bio for presence of keyword

- Follower Wonk - [http://followerwonk.com/](http://followerwonk.com/)

- We now have the seed nodes!
Twitter Bio - Example

Denzil Correa
@denzil_correa
Scientist, Cricketer, Lucid Dreamer, movies.io
Singapore
Expand - Lets grow!

- Get the *strong links* of each seed node

- Each *strong link* is a candidate for the community
Filter - Prune the branches

- Calculate local modularity (R) for each candidate
- Choose the node with the best \( R \)
Iterate - Again & Again ...

- Repeat this until a stopping condition

- Stopping Condition
  - Value of R
  - Size of community
Experimental Setup

- Topics - CAD (Computer Aided Design) and Kashmir

- Search bios with Follower Wonk to obtain seed nodes

- Expand Top \( k = 6, 15, 50, 150 \) strong links

- Stopping Condition - No consistent change in R or negative change
Structural Evaluation
Topic - CAD

CAD: Computer Aided Design

Best Core Community
Nodes = 58, R = 0.7177

- k = 6
- k = 15
- k = 50
- k = 150
Topic - Kashmir

Best Core Community
Nodes = 133, R=0.8298
Community - Kashmir
Benchmark Evaluation

Benchmark-CCD versus iTop on topics: 'CAD' and 'Kashmir'

'CAD' Benchmark Winner: iTop
Nodes = 58, R = 0.7177

'Kashmir' Benchmark Winner: iTop
Nodes = 133, R = 0.8298

Tscherteu and Langreiter [2009]
Semantic Evaluation
Word Cloud - CAD
Word Cloud - Kashmir
## Empirical Evaluation

<table>
<thead>
<tr>
<th>Type</th>
<th>Percentage of users</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>CAD</strong></td>
</tr>
<tr>
<td>Bio contains input topic</td>
<td>17%</td>
</tr>
<tr>
<td>Relevant to input topic</td>
<td>84%</td>
</tr>
<tr>
<td>Members of a topic-relevant <em>list</em></td>
<td>80%</td>
</tr>
</tbody>
</table>
## Network Statistics

<table>
<thead>
<tr>
<th></th>
<th>CAD</th>
<th>Kashmir</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>57</td>
<td>132</td>
</tr>
<tr>
<td><strong>Edges</strong></td>
<td>1292</td>
<td>4444</td>
</tr>
<tr>
<td><strong>Average Path Length</strong></td>
<td>1.7</td>
<td>1.94</td>
</tr>
<tr>
<td><strong>Average Clustering Coefficient</strong></td>
<td>0.712</td>
<td>0.54</td>
</tr>
<tr>
<td><strong>Graph Density</strong></td>
<td>0.405</td>
<td>0.257</td>
</tr>
</tbody>
</table>
Limitations

- Stopping condition as a parameter
- Discovers only core communities, leaves out islands
Research Contributions

- Interaction based and topic-specific community discovery
- Investigation of local information algorithm for community discovery
Conclusion

- Interactions help discover true social communities

- Local information algorithms can be used to discover topic-specific communities

- Evaluation of communities on multiple dimensions shows feasibility
Thank You!

denzilc@iiitd.ac.in or contact me on Twitter @denzil_correa
Simplicity is the ultimate sophistication.