Community Detection in Bipartite Networks: Algorithms and Case Studies

Kathy Horadam and Taher Alzahrani

Mathematical and Geospatial Sciences, RMIT
Melbourne, Australia

IWCNA 2014
Outline

1 Bipartite Networks

2 Community Detection Algorithms

3 Algorithms for Bipartite Networks

4 Our Approach: Apply Infomap Algorithm to Weighted Projection

5 Case Studies

6 Conclusions and Future Work
Bipartite Networks

The network has two node sets $P$ (Primary, of most interest) and $S$ (Secondary). Edges do not occur between nodes in the same set.
Why do we care about bipartiteness?

Many real world examples are naturally bipartite:

- actors and events (in social networks)
- authors and papers (in collaboration networks)
- trains and railway stations (in railway networks)
- companies and goods (in financial networks)

Guillaume & Latapy (2006) [9] argue that any complex network equals a bipartite network through decomposition, and propose a random bipartite graph model (\(\sim\) the configuration model)
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
Aim of Community Detection algorithms: to derive coarse-grain depiction of real large-scale networks

There is a vast number of community detection algorithms available.

Two algorithmic approaches:

- Structural examination of strength of within community connectivity vs between community connectivity.
  - based on underlying stochastic model of network formation
  - e.g. measured by conductance, modularity, link density

- Examining flows across network from which structure/communities emerge which are visited more frequently within than are jumped from.
  - based on how underlying structure constrains flow across network
  - e.g. measured by random walks + teleportation, spectral methods.
Outline

1 Bipartite Networks

2 Community Detection Algorithms

3 Algorithms for Bipartite Networks

4 Our Approach: Apply Infomap Algorithm to Weighted Projection

5 Case Studies

6 Conclusions and Future Work
Performance on benchmark networks

Is there a “best” algorithm?

Public benchmark graph data (Lancichinetti et al [11]) which has:

- Ability to test larger size networks of $10^3$ to $10^5$ nodes.
- Power law distributions for the node degree and community sizes.
- Overlapping communities.
- Directed and weighted graphs.
- Compares partition found by tested algorithm against actual ("planted") partition
Comparison of performance of 12 algorithms [12]

Conclude that “Infomap”, “Louvain” and a Potts model method are the best performing algorithms on these benchmarks.

Infomap and Louvain also very fast, linear in network size, so further tested on benchmark graphs

- with 50,000 and 100,000 nodes
- range of community sizes, from 20 to 1,000 nodes
- maximum degree 200

Performance of Louvain is worse than on smaller graphs, whereas that of Infomap is stable.
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
Modularity-based algorithms: CNM and Louvain

CNM (Clauset-Newman-Moore) algorithm [4]: optimize a quality function modularity $Q$ (of a partition $M$ of a graph with $N$ vertices), OK for graphs up to $N = 10^6$ vertices.

Problem: The resolution limit of modularity [8]

Maximisation of $Q$ will fail to identify communities with edge number $\leq \sqrt{|E|}$ (even when they are cliques).

eg if $m \gg p$, higher modularity for joined pair of cliques than for cliques themselves.

So it may not find important small communities.

Various attempts to overcome this.
“Louvain”, Blondel et.al Fast unfolding algorithm 2008

Multistep technique: local optimization of $Q$ in nbhd of each node [3].

Claims to unfold a complete hierarchical structure and avoid resolution limit problem
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
“Infomap”, Rosvall and Bergstrom 2008, Minimum Description Length [16]

- Minimises a different quality function from $Q$, the “map equation" $L$ of a partition $M$ of the graph.
- $L$ is the average description length of binary codewords describing a single step in a random walk on the graph.
- $L$ is a sum of weighted entropies.

Exploits the duality between compressing information representing a flow on the network and detecting/extracting significant structures in network.
Infomap (Rovall and Bergstrom 2008 [16])
Outline

1 Bipartite Networks

2 Community Detection Algorithms

3 Algorithms for Bipartite Networks

4 Our Approach: Apply Infomap Algorithm to Weighted Projection

5 Case Studies

6 Conclusions and Future Work
Structural algorithms for Bipartite Networks

- Different modularity based algorithms developed by Barber [1], Gumiera [10] and Michel et al [5].
- Label Propagation Algorithm for bipartite network (LPAb) [2]

Table: Numbers of communities of women ($P$) in benchmark Southern women network

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Optimised</th>
<th>Network</th>
<th>Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guimera [10]</td>
<td>modularity</td>
<td>weighted projection</td>
<td>2</td>
</tr>
<tr>
<td>Michel [5]</td>
<td>bimodularity</td>
<td>bipartite</td>
<td>3</td>
</tr>
<tr>
<td>Barber [1]</td>
<td>bimodularity</td>
<td>bipartite</td>
<td>4</td>
</tr>
<tr>
<td>LPAb(+)</td>
<td>bimodularity</td>
<td>bipartite</td>
<td>4</td>
</tr>
</tbody>
</table>
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
Problem: Infomap can’t be applied to a bipartite network

Example

If you start in one set of a bipartite network, then you will always be in that set after an even number of steps, so the probability of being at a particular vertex is zero at odd time steps. In terms of Markov chains, the random walk on a bipartite graph is *periodic*.

The random walk has a stationary distribution on a bipartite graph, but it won’t converge to it. Thus, we can not implement Infomap on bipartite graph because of periodicity.
A different MDL algorithm CAN be applied directly to bipartite networks.

Inference using MDL on stochastic blockmodel (Peixoto 2013 [15]).

BUT

1. There is again a theoretical resolution limit for detection of communities of size \( \leq \sqrt{N} \).
2. Algorithm only tested on small networks.

However, the inferred communities fully reflect the bipartite nature. The communities partition \( P \) and \( S \) separately.

This supports our decision to apply Infomap to the weighted networks projected from \( P \) (or from \( S \)).
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
Method:

- Multiple links - weighted
- self-connections
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
Experiments on five real world bipartite networks

**Table:** Network sizes, where $P$ and $S$ are the number of primary set nodes and secondary set nodes respectively and $m$ is the total number of edges.

<table>
<thead>
<tr>
<th>Network</th>
<th>$P$</th>
<th>$S$</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southern women[6]</td>
<td>18</td>
<td>14</td>
<td>89</td>
</tr>
<tr>
<td>NSW Crimes[14]</td>
<td>155</td>
<td>22</td>
<td>9611</td>
</tr>
<tr>
<td>Noordin Top Terrorists</td>
<td>74</td>
<td>45</td>
<td>276</td>
</tr>
<tr>
<td>Scientific collaboration[13]</td>
<td>16726</td>
<td>22016</td>
<td>58595</td>
</tr>
<tr>
<td>Australian government contracts[7]</td>
<td>11924</td>
<td>1655</td>
<td>70019</td>
</tr>
</tbody>
</table>
Table: Community numbers in $P$, where $L$ is the (minimum) code length and $Q$ is the (maximum) modularity.

<table>
<thead>
<tr>
<th>Network</th>
<th>Infomap</th>
<th>Louvain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comm.</td>
<td>$L$</td>
</tr>
<tr>
<td>Southern women</td>
<td>4</td>
<td>3.992</td>
</tr>
<tr>
<td>NSW crime</td>
<td>2</td>
<td>7.276</td>
</tr>
<tr>
<td>Noordin Top Terrorists</td>
<td>5</td>
<td>5.846</td>
</tr>
<tr>
<td>Australian government</td>
<td>1114</td>
<td>8.340</td>
</tr>
<tr>
<td>Scientific collaboration</td>
<td>2131</td>
<td>6.164</td>
</tr>
</tbody>
</table>
Outline

1. Bipartite Networks

2. Community Detection Algorithms

3. Algorithms for Bipartite Networks

4. Our Approach: Apply Infomap Algorithm to Weighted Projection

5. Case Studies

6. Conclusions and Future Work
Southern women network [6] (women $P$ and events $S$)

The four communities of women found in the Southern women dataset. Red nodes represent $S$, the events the women attended, and the four other colors represent four communities within $P$, with nodes labelled by first name. The 2 women in each smallest community are core members of their social networks.
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
NSW crimes (LGAs P and crimes S)
Outline

1. Bipartite Networks
2. Community Detection Algorithms
3. Algorithms for Bipartite Networks
4. Our Approach: Apply Infomap Algorithm to Weighted Projection
5. Case Studies
6. Conclusions and Future Work
The two small cliques we find are the Bali II (2005) bombers (Community 5) and Ring Baten members involved in bombing the Australian embassy in 2004 (Community 4). In Community 1, 17 of the 25 members belonged to Jemaah Islamiyah, a transnational Southeast Asian militant Islamist terrorist organisation linked to Al-Qaeda.
Conclusions and future work

- Integration of projection with Infomap results in more valuable information about small strong communities than the high performance modularity based algorithm Louvain.
Conclusions and future work

- Integration of projection with Infomap results in more valuable information about small strong communities than the high performance modularity based algorithm Louvain.
- Neither Louvain nor Infomap finds overlapping communities. We may merge communities found separately in $P$ and $S$ to do this and recover communities in the bipartite graph.

Acknowledgements: Thanks to Serdar Boztas, Chris Bellman, Sarah Taylor and Murray Aitken. The speaker is partly supported by Department of Defence of Australia Agreement 4500743680.
Conclusions and Future Work

Conclusions and future work

- Integration of projection with Infomap results in more valuable information about small strong communities than the high performance modularity based algorithm Louvain.
- Neither Louvain nor Infomap finds overlapping communities. We may merge communities found separately in $P$ and $S$ to do this and recover communities in the bipartite graph.
- We will compare the communities found this way with those found by modularity and message-length algorithms using eg NMI (normalised mutual information).
Conclusions and future work

- Integration of projection with Infomap results in more valuable information about small strong communities than the high performance modularity based algorithm Louvain.
- Neither Louvain nor Infomap finds overlapping communities. We may merge communities found separately in $P$ and $S$ to do this and recover communities in the bipartite graph.
- We will compare the communities found this way with those found by modularity and message-length algorithms using eg NMI (normalised mutual information).

Acknowledgements: Thanks to Serdar Boztas, Chris Bellman, Sarah Taylor and Murray Aitken. The speaker is partly supported by Department of Defence of Australia Agreement 4500743680.


