Learning Community Structure in Networks with Metadata

Aaron Clauset
@aaronclauset
Computer Science Dept. & BioFrontiers Institute
University of Colorado, Boulder
External Faculty, Santa Fe Institute
what is community structure?

vertices with same pattern of inter-community connections

community detection:

community interaction matrix
metadata vs. community detection
metadata vs. community detection

many networks include *metadata* on their nodes:

- social networks: age, sex, ethnicity or race, etc.
- food webs: feeding mode, species body mass, etc.
- Internet: data capacity, physical location, etc.
- protein interactions: molecular weight, association with cancer, etc.

metadata $\mathbf{x}$ is often used to evaluate the accuracy of community detection algs.

if community detection method $\mathcal{A}$ finds a partition $\mathcal{P}$ that correlates with $\mathbf{x}$ then we say that $\mathcal{A}$ is good
metadata vs. community detection

NCAA 2000 schedule

political blogs network
metadata vs. community detection

often, groups found by community detection are meaningful

• allegiances or personal interests in social networks [1]
• biological function in metabolic networks [2]

but

[1] see Fortunato (2010), and Adamic & Glance (2005)
metadata vs. community detection

often, groups found by community detection are meaningful

• allegiances or personal interests in social networks [1]
• biological function in metabolic networks [2]

but some recent studies claim these are the exception

• real networks **either** do not contain structural communities **or**
  communities exist but they do not correlate with metadata groups [3]

[1] see Fortunato (2010), and Adamic & Glance (2005)
[3] see Leskovec et al. (2009), and Yang & Leskovec (2012), and Hric, Darst & Fortunato (2014)
Hric, Darst & Fortunato (2014)

- 115 networks with metadata & 12 community detection methods
- compare extracted $P$ with observed $x$ for each $A$

<table>
<thead>
<tr>
<th>Name</th>
<th>No. Nodes</th>
<th>No. Edges</th>
<th>No. Groups</th>
<th>Description of group nature</th>
</tr>
</thead>
<tbody>
<tr>
<td>lfr</td>
<td>1000</td>
<td>9839</td>
<td>40</td>
<td>artificial network (lfr, 1000S, $\mu = 0.5$)</td>
</tr>
<tr>
<td>karate</td>
<td>34</td>
<td>78</td>
<td>2</td>
<td>membership after the split</td>
</tr>
<tr>
<td>football</td>
<td>115</td>
<td>615</td>
<td>12</td>
<td>team scheduling groups</td>
</tr>
<tr>
<td>polbooks</td>
<td>105</td>
<td>441</td>
<td>2</td>
<td>political alignment</td>
</tr>
<tr>
<td>polblogs</td>
<td>1222</td>
<td>16782</td>
<td>3</td>
<td>political alignment</td>
</tr>
<tr>
<td>dpd</td>
<td>35029</td>
<td>161313</td>
<td>580</td>
<td>software package categories</td>
</tr>
<tr>
<td>as-caida</td>
<td>46676</td>
<td>262953</td>
<td>225</td>
<td>countries</td>
</tr>
<tr>
<td>fb100</td>
<td>762–41536</td>
<td>16651–146564</td>
<td>2–2597</td>
<td>common students’ traits</td>
</tr>
<tr>
<td>pgp</td>
<td>81036</td>
<td>190143</td>
<td>17824</td>
<td>email domains</td>
</tr>
<tr>
<td>anobii</td>
<td>136547</td>
<td>892377</td>
<td>25992</td>
<td>declared group membership</td>
</tr>
<tr>
<td>dblp</td>
<td>317080</td>
<td>1049866</td>
<td>13472</td>
<td>publication venues</td>
</tr>
<tr>
<td>amazon</td>
<td>366997</td>
<td>1231439</td>
<td>14–29432</td>
<td>product categories</td>
</tr>
<tr>
<td>flickr</td>
<td>1715255</td>
<td>22613981</td>
<td>101192</td>
<td>declared group membership</td>
</tr>
<tr>
<td>orkut</td>
<td>3072441</td>
<td>117185083</td>
<td>8730807</td>
<td>declared group membership</td>
</tr>
<tr>
<td>lj-backstrom</td>
<td>4843953</td>
<td>43362750</td>
<td>292222</td>
<td>declared group membership</td>
</tr>
<tr>
<td>lj-mislove</td>
<td>5189809</td>
<td>49151786</td>
<td>2183754</td>
<td>declared group membership</td>
</tr>
</tbody>
</table>

[1] fb100 is 100 networks
metadata vs. community detection

Hric, Darst & Fortunato (2014)

• evaluate by normalized mutual information $\text{NMI}(\mathcal{P}, \mathbf{x})$

---

Figure 5: (Color online) NMI scores between structural communities and metadata groups for different networks. Scores are grouped by datasets on the $x$-axis. The height of each column is the maximal NMI score between any partition layer of the metadata partitions and any layer returned by the community detection method, considering only those comparisons where the overlap of the partitions is larger than 10% of total number of nodes.

Methods do not align with partitions built from metadata, but what about specific groups? Can we detect any of the groups well? Are some groups reflected in the graph structure and detectable, but lost in the bulk noise of the graph? This is what we wish to investigate here.

The basis of our analysis is the Jaccard score between two groups. Let $C_i$ represent (the set of nodes of) the known group $i$, and $D_j$ represent (the set of nodes of) the detected community $j$. The Jaccard score between these two sets is defined as

$$J(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i | + |D_j| - |C_i \cap D_j|},$$

with $|· · ·|$ set cardinality, $\cap$ set intersection, and $\cup$ set union. The Jaccard score ranges from one (perfect match) to zero and roughly indicates the fraction of nodes shared between the two sets: the match quality.

The recall score measures how well one known group is detected. The recall score of one known group $C_i$ is defined as the maximal Jaccard score between it and every detected community $D_j$,

$$\text{R}(C_i) = \max_{D_j \in \{D_j\}} J(C_i, D_j).$$

It is near one if the group is well detected and low otherwise. We can study the distribution of these scores to see how many groups can be detected at any given quality level. Recall measures the detection of known groups, and to measure the significance of detected communities, we can reverse the measure to calculate a precision score $P(D_j) = \max_{C_i \in \{C_i\}} J(D_j, C_i)$.

The precision score tells us how well one detected community corresponds to any known group.

We can now directly quantify the two conditions for good community detection: every known group must correspond to some detected community, and every detected community must represent some known group. Both of these measures are still interesting independently: a high recall but low precision indicates that the known groups are reflected in the network structurally, but there are many structural communities that are not known. We visualize the scores by means of rank-Jaccard plots which give an overview of the network's detection quality. We compute the recall (precision) for every known (detected) group and sort the groups in order of ascending Jaccard score. We plot recall (precision) vs the group rank, sorted by recall (precision) score so that the horizontal scale is the relative group rank, i.e. the ratio between the rank of the group and the number of groups (yielding a value between 0 and 1). Similar to our treatment of the partition-level analysis, we only plot matchings whose intersection covers more than 10% of total nodes in the graph. In our final plots, the average value of the curve (proportional to the area under it) is the average recall or precision score over all groups. The shape of the curve can tell us if all groups are detected equally well (yielding a high plateau) or if there is a large inequality in detection (a high slope). Furthermore, this allows us to compactly represent multiple layers. Each independent layer of known (detected) groups can be plotted in the same figure. We would generally look for the highest curve to know if any layer has a high recall (precision). When computing recall (precision), unless otherwise specified, as detected communities we consider the communities of all partitions delivered by a method, whereas the metadata groups are those present in all metadata partitions (if more than one partition is available in either case). This will give us the maximum possible recall (precision), which might be far higher than $\text{NMI}(\mathcal{P}, \mathbf{x})$. [1]

---

[1] maximum NMI between any partition layer of the metadata partitions and any layer returned by the community detection method
what is the goal of community detection?

network $G$ + method $f \rightarrow$ communities $C = f(G)$ vs. $M$ metadata
the ground truth about metadata

what is the goal of community detection?

network $G$ + method $f \rightarrow$ communities $C = f(G)$ vs. $M$ metadata

$C \approx M$

"this method works!"
the ground truth about metadata

what is the goal of community detection?

network $G$ + method $f \rightarrow$ communities $C = f(G)$ vs. $M$ metadata

C $\approx$ M

"this method works!"

C $\neq$ M

"this method stinks!"
what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata $M$ are unrelated to network structure $G$
the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find \( f(G) = C \neq M \):

1. metadata \( M \) are unrelated to network structure \( G \)
2. metadata \( M \) and communities \( C \) capture different aspects of structure

social groups

leaders and followers
the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find $f(G) = C \neq M$:

1. metadata $M$ are unrelated to network structure $G$
2. metadata $M$ and communities $C$ capture different aspects of structure
3. network $G$ has no community structure
the ground truth about metadata

what is the goal of community detection?

there are 4 indistinguishable reasons why we might find \( f(G) = C \neq M \):

1. metadata \( M \) are unrelated to network structure \( G \)
2. metadata \( M \) and communities \( C \) capture different aspects of structure
3. network \( G \) has no community structure
4. algorithm \( f \) is bad

"this method stinks!"
theorems for community detection

1. Theorem: no bijection between ground truth and communities

\[ g_1(T_1) \rightarrow G \]
\[ g_2(T_2) \rightarrow G \]

different processes with different ground truths but same observed network

2. Theorem: No Free Lunch in community detection

no algorithm \( f \) has better performance than any other algorithm \( f' \), averaged over all possible inputs \( \{G\} \)

\[ \rightarrow \text{good performance comes from matching algorithm } f \text{ to its preferred subclass of networks } \{G'\} \subseteq \{G\} \]

[1] performance defined as adjusted mutual information (AMI), which is like the normalized mutual information, but adjusted for expected values
[3] proofs of these theorems is in Peel, Larremore, Clauset (2017)
Theorem: *there is no ground truth in real-world data*

DON'T TRY TO FIND THE GROUND TRUTH

INSTEAD . . . TRY TO REALIZE THERE IS NO GROUND TRUTH
but wait!

[1] image copyright BostonGazette or maybe 20th Century Fox? gah
idea:

use metadata $\mathbf{x}$ to help select a partition $\mathcal{P}^* \in \{\mathcal{P}\}$ that correlates with $\mathbf{x}$, from among the exponential number of plausible partitions.
idea:

use metadata $\mathbf{x}$ to help select a partition $\mathcal{P}^* \in \{\mathcal{P}\}$ that correlates with $\mathbf{x}$, from among the exponential number of plausible partitions

use a generative model to guide the selection:

• define a parametric probability distribution over networks $\Pr(G | \theta)$
• generation: given $\theta$, draw $G$ from this distribution
• inference: given $G$, choose $\theta$ that makes $G$ likely

machine learning to the rescue!
the stochastic block model

- each vertex $u$ has type $s_u \in \{1, \ldots, k\}$
- stochastic block matrix $\Theta$ of group-level connection probabilities
- probability that $P(A_{uv} = 1) = \theta_{s_u, s_v}$

community = vertices with same pattern of inter-community connections
a metadata-aware stochastic block model

generation

given metadata \( \mathbf{x} = \{ x_u \} \) and degree \( \mathbf{d} = \{ d_u \} \) for each node \( u \)
- each node \( u \) is assigned a community \( s \) with probability \( \gamma_{sx} \)
- thus, prior on community assignments is \( P(s | \Gamma, \mathbf{x}) = \prod \gamma_{s_i, x_i} \)
- given assignments, place edges independently, each with probability:
  \[
p_{uv} = d_u d_v \theta_{s_u, s_v}
\]
- where the \( \theta_{st} \) are the stochastic block matrix parameters

this is a degree-corrected stochastic block model (DC-SBM)

with a metadata-based prior on community labels

[1] \( \Gamma \) is the \( k \times K \) matrix of parameters \( \gamma_{sx} \)
a metadata-aware stochastic block model

inference

given observed network \( A \) (adjacency matrix)

- the model likelihood is

\[
P(A \mid \Theta, \Gamma, x) = \sum_s P(A \mid \Theta, s) P(s \mid \Gamma, x)
\]

\[
= \sum_s \prod_{u < v} p_{uv}^A (1 - p_{uv})^{1-A_{uv}} \prod_u \gamma_{s_u, x_u}
\]

- where \( \Theta \) is a \( k \times k \) matrix of community interaction parameters \( \theta_{st} \), and the sum is over all possible assignments \( s \)

- we fit this model to data using expectation-maximization (EM) to maximize \( P(A \mid \Theta, \Gamma, x) \) w.r.t. \( \Theta \) and \( \Gamma \)

networks with planted structure

does this method recover known structure in synthetic data?
does this method recover known structure in synthetic data?

- use SBM to generate planted partition networks, with $k = 2$ equal-sized groups and mean degree $c = (c_{in} + c_{out})/2$
- assign metadata with variable correlation $\rho \in [0.5, 0.9]$ to true group labels
- vary strength of partition $c_{in} - c_{out}$
- when $c_{in} - c_{out} \leq \sqrt{2(c_{in} + c_{out})}$, no structure-only algorithm can recover the planted communities better than chance (the detectability threshold, which is a phase transition)

networks with planted structure

let mean degree $c = 8$

- when $\rho = 0.5$, metadata isn't useful and we recover regular SBM behavior
networks with planted structure

let mean degree $c = 8$

• when $\rho = 0.5$, metadata isn’t useful and we recover regular SBM behavior

• when metadata correlates with true groups, $\rho > 0.5$, accuracy is better than either metadata or SBM alone

(metadata + SBM performs better than either

• any algorithm without metadata, or

• metadata alone.)
real-world networks
real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

2. **marine food web**: predator-prey interactions among 488 species in Weddell Sea in Antarctica

3. **Malaria gene recombinations**: recombination events among 297 var genes

4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni

5. **Internet graph**: peering relations among 46,676 Autonomous Systems
real-world networks

1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

   - $\mathbf{x} = \{\text{grade 7-12, ethnicity, gender}\}$

[1] Add Health network data, designed by Udry, Bearman & Harris
1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

   - \( x = \{\text{grade 7-12, ethnicity, gender}\} \)
   - method finds a good partition between high-school and middle-school
     \[ \text{NMI} = 0.881 \]
   - without metadata:
     \[ \text{NMI} \in [0.105, 0.384] \]

[1] Add Health network data, designed by Udry, Bearman & Harris
1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

- \( x = \{\text{grade 7-12, ethnicity, gender}\} \)

- method finds a good partition between blacks and whites (with others scattered among)

  \[ \text{NMI} = 0.820 \]

- without metadata:
  \[ \text{NMI} \in [0.120, 0.239] \]

[1] Add Health network data, designed by Udry, Bearman & Harris
1. **high school social network**: 795 students in a medium-sized American high school and its feeder middle school

- $x = \{\text{grade 7-12, ethnicity, gender}\}$

- method finds no good partition between males/females. instead, chooses a mixture of grade/ethnicity partitions

  \[
  \text{NMI} = 0.003
  \]

- without metadata:

  \[
  \text{NMI} \in [0.000, 0.010]
  \]

[1] Add Health network data, designed by Udry, Bearman & Harris
parting thoughts
a metadata-aware stochastic block model

• probabilistic model of community structure and node metadata

$$P(A | \Theta, \Gamma, x)$$

• **yields**: posterior probabilities of community labels $q$, the mixing matrix $\Theta$, and learned metadata-community association $\Gamma$

• metaSBM performs better than any either structure-only or metadata-only algorithm

• highly scalable, via EM + belief propagation

• works well in practice
metadata is just more data

• there is no ground truth in real-world networks (only more data)

• many reasons an algorithm may fail at recovering metadata partition

• # of good partitions of $G$ grows exponentially with $n$

• solution: use metadata $\mathbf{x}$ to guide the choice to a partition that correlates with metadata $\mathbf{x}$ (if the exist)

• solution: make community detection supervised learning

• but be careful, there is No Free Lunch
Structure and inference in annotated networks

M. E. J. Newman\(^1,\)\(^2\) and Aaron Clauset\(^2,\)\(^3\)

\(^1\)Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA
\(^2\)Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501
\(^3\)Department of Computer Science and BioFrontiers Institute, University of Colorado, Boulder, CO 80309, USA

The ground truth about metadata and community detection in networks

Leto Peel\(^1,\)\(^2,\)\(^\ast\), Daniel B. Larremore\(^3,\)\(^\dagger\), and Aaron Clauset\(^4,\)\(^5,\)\(^3,\)\(^\ddagger\)

\(^1\)ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, Belgium
\(^2\)naXys, Université de Namur, Namur, Belgium
\(^3\)Santa Fe Institute, Santa Fe, NM 87501, USA
\(^4\)Department of Computer Science, University of Colorado, Boulder, CO 80309, USA
\(^5\)BioFrontiers Institute, University of Colorado, Boulder, CO 80309, USA

\(\ast\) Corresponding author.
\(\dagger\) These authors contributed equally.

Acknowledgements

Mark Newman (Michigan)

Leto Peel (Louvain)
Daniel B. Larremore (Santa Fe)

calculate alignment scores
convert to alignment indicators
remove short aligned regions
extract highly variable regions

community detection
2. **marine food web**: predator-prey interactions among 488 species in Weddell Sea in Antarctica

- **x** = \{**species body mass**, feeding mode, oceanic zone\}

- partition recovers known correlation between body mass, trophic level, and ecosystem role:

![Graph showing probability of community membership vs mean body mass](image)

\[
\begin{align*}
10^{-12} & \quad 10^{-9} & \quad 10^{-6} & \quad 10^{-3} & \quad 10^{0} & \quad 10^{3} & \quad 10^{6} & \quad 10^{9} \\
0 & \quad 0.5 & \quad 1
\end{align*}
\]

The communities in the HVR 6 network represent highly non-random patterns of recombination, which are thought to indicate functional constraints on protein structure. Previous work has conjectured that common constraints on recombination span distinct HVRs. We can test this hypothesis using the methods described in this paper. There is no reason a priori to expect that the community structure of HVR 6 should correlate with that of HVR 5 because the Cys and CP labels are derived from outside the HVR 5 sequences—Cys labels reflect cysteine counts in HVR 6 while CP labels subdivide Cys labels based on sequence motifs adjacent to, but outside of, HVR 5.
3. **Malaria gene recombinations**: recombination events among 297 var genes

- $x = \{Cys$-PoLV labels for HVR6 region\}$
- with metadata, partition discovers correlation with Cys labels (which are associated with severe disease)

\[
\text{HVR6}
\]

without metadata

\[\text{NMI} \in [0.077, 0.675]\]

with metadata

\[\text{NMI} = 0.596\]

3. **Malaria gene recombinations**: recombination events among 297 *var* genes

- \( \mathbf{x} = \{ \text{Cys-PoLV labels for HVR6 region} \} \)

- on adjacent region of gene, we find Cys-PoLV labels correlate with recombinant structure here, too

---

real-world networks

4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

- \( x = \{ \text{graduation year}, \text{dormitory} \} \)
- method finds a good partition between alumni, recent graduates, upperclassmen, sophomores, and freshmen

\[ \text{NMI} = 0.668 \]

- without metadata:
  \[ \text{NMI} \in [0.573, 0.641] \]
4. **Facebook friendships**: online friendships among 15,126 Harvard students and alumni (in Sept. 2005)

- \( x = \{ \text{graduation year, dormitory} \} \)
- method finds a good partition among the dorms
- \( \text{NMI} = 0.255 \)
- without metadata:
  \( \text{NMI} \in [0.074, 0.224] \)
real-world networks

5. **Internet graph**: 262,953 peering relations among 46,676 Autonomous Systems

- \( x = \{\text{country location of AS}\} \)
- method finds a good partition along the lines of the 173 countries

\[
NMI = 0.870
\]

- without metadata:

\[
NMI \in [0.398, 0.626]
\]