Community structure: evaluation and motif analysis in link streams

Jean Creusefond, GREYC, Normandy University
Work with: Sylvain Peyronnet, Thomas Largillier, Remy Cazabet
LIP6 Presentation

May 12, 2016
Ground-truths and quality functions
Social network:
Community detection and quality functions

A clustering:
Community detection and quality functions

Two clusterings:

How to chose the best?

We use quality functions, for optimisation and evaluation.
Algorithms application

\[ \text{Alg1} \times \text{Graph1} \times \text{Graph2} \Rightarrow \text{Alg1} \times \text{Graph1} \times \text{Graph2} \]
## Comparison with ground-truth

<table>
<thead>
<tr>
<th>Alg1</th>
<th>Graph1</th>
<th>Alg2</th>
<th>Graph2</th>
<th><strong>Alg3</strong></th>
<th>Graph2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="Graph1" /></td>
<td><img src="image2" alt="Graph1" /></td>
<td><img src="image1" alt="Graph2" /></td>
<td><img src="image2" alt="Graph2" /></td>
<td><img src="image3" alt="Graph2" /></td>
</tr>
<tr>
<td><img src="image4" alt="Alg1" /></td>
<td><img src="image5" alt="Alg1" /></td>
<td><img src="image6" alt="Alg2" /></td>
<td><img src="image7" alt="Alg2" /></td>
<td><img src="image8" alt="Alg3" /></td>
<td><img src="image9" alt="Alg3" /></td>
</tr>
<tr>
<td><img src="image10" alt="Alg1" /></td>
<td><img src="image11" alt="Alg1" /></td>
<td><img src="image12" alt="Alg1" /></td>
<td><img src="image13" alt="Alg1" /></td>
<td><img src="image14" alt="Alg2" /></td>
<td><img src="image15" alt="Alg2" /></td>
</tr>
<tr>
<td><img src="image16" alt="Alg1" /></td>
<td><img src="image17" alt="Alg1" /></td>
<td><img src="image18" alt="Alg1" /></td>
<td><img src="image19" alt="Alg1" /></td>
<td><img src="image20" alt="Alg3" /></td>
<td><img src="image21" alt="Alg3" /></td>
</tr>
</tbody>
</table>

- **Ground-truth1**: ![Ground-truth](image22)
- **Ground-truth2**: ![Ground-truth](image23)

<table>
<thead>
<tr>
<th>Alg1</th>
<th><strong>Graph1</strong></th>
<th>Alg2</th>
<th><strong>Graph2</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image24" alt="Alg1" /></td>
<td>gold-standard : ...</td>
<td><img src="image25" alt="Alg2" /></td>
<td>trustable : ...</td>
</tr>
<tr>
<td><img src="image26" alt="Alg1" /></td>
<td>value : ...</td>
<td><img src="image27" alt="Alg3" /></td>
<td>...</td>
</tr>
<tr>
<td><img src="image28" alt="Alg3" /></td>
<td>...</td>
<td><img src="image29" alt="Alg3" /></td>
<td>...</td>
</tr>
</tbody>
</table>

Jean Creusefond, GREYC, Normandy University | Quality and motifs | 5/25
Application of quality functions

<table>
<thead>
<tr>
<th></th>
<th>Graph1</th>
<th>Graph2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg1</td>
<td><img src="Alg1" alt="Graph1" /></td>
<td><img src="Alg1" alt="Graph2" /></td>
</tr>
<tr>
<td>Alg2</td>
<td><img src="Alg2" alt="Graph1" /></td>
<td><img src="Alg2" alt="Graph2" /></td>
</tr>
<tr>
<td>Alg3</td>
<td><img src="Alg3" alt="Graph1" /></td>
<td><img src="Alg3" alt="Graph2" /></td>
</tr>
</tbody>
</table>

Quality1 × Quality2 \(\Rightarrow\) Quality3

<table>
<thead>
<tr>
<th></th>
<th>Quality1</th>
<th>Quality2</th>
<th>Quality3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Graph1</td>
<td>Graph2</td>
<td>Graph1</td>
</tr>
<tr>
<td>Alg1</td>
<td>quality score</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Alg2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Alg3</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Coherence quantification

<table>
<thead>
<tr>
<th></th>
<th>Quality1</th>
<th>Quality2</th>
<th>Quality3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Graph1</td>
<td>Graph2</td>
<td>Graph1</td>
</tr>
<tr>
<td>Alg1</td>
<td>quality score</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Alg2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Alg3</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Graph1</th>
<th>Graph2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg1</td>
<td>gold-standard</td>
<td>...</td>
</tr>
<tr>
<td>Alg2</td>
<td>trustable</td>
<td>...</td>
</tr>
<tr>
<td>Alg3</td>
<td>value</td>
<td>...</td>
</tr>
</tbody>
</table>

- red: good coherence
- blue: bad coherence
We measure the coherence of the quality functions with ground-truth data.
Context: ground-truths where quality functions behave the same way.

<table>
<thead>
<tr>
<th>Quality1</th>
<th>Quality2</th>
<th>Quality3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Graph2</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Graph1 $\Rightarrow$ Graph2

Jean Creusefond, GREYC, Normandy University
Need for data

- **Ground-truths**: Communities that can be trusted
- **Algorithms**: They should have various designs.
- **Quality functions**: The main items to compare
- **Comparison methods**: Multiple functions output complementary results
### Normalised Mutual Information (NMI)

Captures the quantity of information needed to infer one clustering from the other.

### F-BCubed

The average ratio, over all individuals, of neighbors in one clustering that are still neighbors in the other one.
Is the methodology able to recognise networks with a similar structure? LFR benchmark: tunable virtual graphs, with social-network structure.

Results

- NMI: Globally coherent with our expectations, but influenced by random generation
- F-BCubed: More robust, difference of overlapping over-matches

Table 1: NMI

Table 2: F-BCubed
(flickr, lj, youtube) : Online social networks

<table>
<thead>
<tr>
<th>file\quality</th>
<th>cc</th>
<th>fb3</th>
<th>mod</th>
<th>nmi</th>
<th>perm</th>
<th>sign</th>
<th>cond</th>
<th>FOMD</th>
<th>comp</th>
<th>cut_ratio</th>
<th>f-odf</th>
<th>sur</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>actors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amazon</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cora</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dblp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flickr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>football</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>github</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lj</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>youtube</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(flickr, lj, youtube) : Modularity
A community detection tool

Efficient community detection

Ground-truth analysis

Plug and play

Extension

Visualisation

Modularity

A community detection tool

CoDADCom (Community Detection Algorithm Comparator) is a software which purpose is to simplify the research in community detection. It is designed to:

- Run multiple community detection algorithms on graphs
- Output a set of statistics on these graphs and on the runs, from degree distribution to the quality of the different communities
- Take into account ground-truth by comparing it with the results of community detection
- Include home-made implementations of (potentially new) community detection algorithms with no code re-writing
- Include user-written quality functions
- Include user-written extrinsic comparison functions

Efficient community detection

The community detection implementations featured in CoDADCom are, for the most part, written in C/C++. For the fastest (loglinear), it is entirely possible to analyse graphs with millions of edges on a standard desktop in less than an hour. Using a server grants the advantage of running all the couples (methods,
Communities and temporal motifs
Definition

Hypothesis

The communication structure (i.e. frequent motifs) is different inside and outside communities.

A **motif** is a regularly repeated communication pattern. A motif has a depth (distance from origin), a size (number of nodes) and a level (number of edges).

![Motifs](image)

**Figure 1**: All level 2 motifs
Traditionally (Zhao et al. 2010, Tabourier et al. 2012), causality is assessed if there is a small time (parameter $W$) difference between emission of the messages.

\[
\begin{array}{c}
0 \\
1 \\
1000 \\
\alpha \quad \beta \quad \gamma \\
\alpha \quad \beta \quad \gamma \\
\alpha \quad \beta \quad \gamma
\end{array}
\]

Not adapted to asynchronous communication.

In this example, AB-AC-CA is not a motif if $W < 999$.

What if $C$ was just away?
Activity periods

Messages received \hspace{2cm} Messages sent

- Activity period 1
- Activity period 2

\text{Jean Creusefond, GREYC, Normandy University} \\
\text{Qualit} \overset{\text{y and motifs}}{\text{y and motifs}}} 18/25
Example of activity period-based motifs

\[ \alpha \beta - \beta \gamma - \alpha \beta \]
\[ AB-BC-AB \]
\[ \alpha \beta - \beta \gamma - \gamma \alpha \]
\[ AB-BC-CA \]
\[ \beta \gamma - \gamma \alpha - \alpha \gamma \]
\[ AB-BC-CB \]
A null model, generated from a base graph, is the same except for a structural property that has been randomised.

**Objective**
By differentiating the null model and the base graph, one can isolate the influence of the randomised property.

**Time-mixing model**: the timestamps of communications are shuffled for each user \(\Rightarrow\) causality is destroyed
Assumption: the measured values follow a gaussian distribution in the null model (checked in practice).
Therefore:

- low ($\sim 0.3\%$) probability that a value from this distribution would be far from the average ($3 \times \sigma$)
- a point far from the average probably does not come from this distribution
- the distribution difference is due to the destruction of causality (the only property that is randomised)
Experiments : data

Data from KONECT (http://konect.uni-koblenz.de/):
- Digg : reply-to
- lkml : reply-to
- slashdot : reply-to
- radoslaw : mail
- Enron : mail
- Facebook wall : post-to
- UC Irvine : instant message

Memberships : iLCD, Louvain and infomap (on aggregated networks)
We also have access to a reply-to network with an overlapping group structure (threads):
- Debian : reply-to (with membership)
Experiments: temporal triangles (1)
Experiments: temporal triangles (2)

Triangles are more frequent and short than the null model $\Rightarrow$ we are detecting structure.
Slightly less included in communities, multiple possible explanation:
- Individuals inside of communities use various means of communication
- Peripheral communications need structure
No structure inside communities?
No pattern has a surprisingly high community score, except with debian membership.
Negociation with my university to get anonymised data about mails and users.

Using motifs for user categorisation
Analysis at user level : if a user emits a lot of some patterns, does that imply a role for him?