Inference and Validation of Networks
ECML/PKDD 2009

Ilias N. Flaounas\textsuperscript{1}, Marco Turchi\textsuperscript{2},
Tijl De Bie\textsuperscript{2} and Nello Cristianini\textsuperscript{1,2}

Department of Computer Science, Bristol University\textsuperscript{1}
Department of Engineering Mathematics, Bristol University\textsuperscript{2}

September 10, 2009
Some Network Inference Examples

Karate Club (Zachary, 1972)

Emails (Adamic et al, 2003)

Dating (Berman et al, 2004)

Internet Nodes (Alvarez et al, 2007)
Some more difficult examples

Gene pathways
(Conti et al, 2007)

Protein interactions
(Jeong et al, 2001)
Our problem: The News Media Network

But, is this network valid? How to validate a network inferred from data?
Pattern Validation

In general, patterns found in data can be validated in two different ways:

- by assessing their **significance**
- by measuring their **predictive power**
Network Validation

We argue that inferred network need to satisfy some key properties:

- It should to be stable.
- It should be related to any available ground truth.
- The topology should be predictable.

The first two properties can be verified by testing if the inferred network is similar to a reference network - **Hypothesis Testing**. We verify the third property by using **Generalised Linear Models**.
Hypothesis Testing

The key idea is to quantify the probability that a test statistic evaluated on observed data could have been found also in random data ($p$-value).

The $p$-value is calculated by:

\[ p = P_{G \sim H_0}(t_{G_R}(G) \leq t_{G_R}(G_I)) \]  

\[ p \approx \frac{\#\{G : t_{G_R}(G) \leq t_{G_R}(G_I)\} + 1}{K + 1} \]

A significance threshold $\alpha$ is chosen, and the null hypothesis is rejected if $p < \alpha$. 

As a test statistic we will use the similarity or distance of the inferred network to a reference network. In literature several approaches have been proposed for the measurement of similarity between networks. In the case of comparing two networks $G_A$ and $G_B$ that have the same set of nodes, we can use Jaccard Distance.

**Jaccard Distance**

$$JD(E_A, E_B) = \frac{|E_A \cup E_B| - |E_A \cap E_B|}{|E_A \cup E_B|}$$  \hspace{1cm} (3)

A value of zero indicates identical networks, and a value of one indicates no shared edges.
Null Models

In the case of validating network patterns, we need to specify a model of random network generation.

**Erdös-Rényi Model**
- A random graph $G(n, p)$ is comprised of $n$ nodes
- Two nodes have a probability $p$ to be connected

**Switching Randomisation**
Starts from a given graph and randomises it by switching edges:

![Switching Randomisation Diagram](image)
Prediction of Topology

If one or more ground truth components are available, they are expected to be able to ‘explain’ the existence of some edges of the network.

- The combination of ground truth elements can be made using Generalized Linear Models (GLMs).
- To assess the ability of GLM to predict the network structure we use a methodology similar to this found in supervised classification.
- Area Under Curve (AUC) was used as the performance measure.
The News Outlets case

We analysed a dataset of
- 1,017,348 news articles
- over a period of 12 consecutive weeks (from October 1st, 2008)
- 543 news outlets
- 32 different locations
- 7 different media types (e.g., newspapers, blogs, etc)
- 22 different languages
- 81,816 stories (clusters of articles)
Comparison of three Inference methods

- **Method A**: Connection of two outlets if they share a minimum number of stories.
- **Method B**: Based on TF-IDF scheme where outlets correspond to documents and clusters to terms.
- **Method C**: Use a weighting scheme for each cluster, based on how common it is among outlets.
Validation of Outlets Network

First property states that the network should be **stable**.

- The reference network would be a network inferred based on independent data
- We use 12 different datasets, one for each of the 12 weeks
- $K = 1000$ random networks per null model
We measure the distance of the network inferred from data of Week 1, to the network inferred from data of the previous week (Week 0).
Stability

We measure the distance of the network inferred from data of Week 1, to a random network constructed based on the Erdős Model.
Stability

We measure the distance of the network inferred from data of Week 1, to a random network constructed based on the Switching Model.
Stability

We measure the distance of the network inferred from data of Week 2, to the network inferred from data of the previous week (Week 1).
We measure the distance of the network inferred from data of Week 2, to the networks inferred from the Null models.
Stability

Each week has three comparisons:

- to the network of the previous week.
- to an Erdös Random Graph
- to a Switching Randomized Graph
Stability

Each week has three comparisons:

- to the network of the previous week.
- to an Erdös Random Graph
- to a Switching Randomized Graph
Stability

We compare the inferred networks of 12 sequential weeks.

The networks inferred by each method are significantly similar ($p < 0.001$) to those inferred the previous week with the same method, under both null models.
Comparison to ground truth components

We compare to the corresponding networks of some available ground truth networks. For the case of news outlets we compared to the networks of:

- Location
- Language
- Type

Example

The ‘Location’ network is comprised of cliques:
Each week has three comparisons:

- to the real Location-Network
- to the corresponding Erdős Random Graph
- to the corresponding Switching Randomized Graph
Language

Each week has three comparisons:

- to the real Language-Network
- to the corresponding Erdős Random Graph
- to the corresponding Switching Randomized Graph
Media Type

Each week has three comparisons:
- to the real Media Type-Network
- to the corresponding Erdös Random Graph
- to the corresponding Switching Randomized Graph
Selecting Inference Method

The selection of the inference method will be made based on their ability to create significant results.

Table: The number of weeks that each Method presented significant results with $p < 0.001$.

<table>
<thead>
<tr>
<th></th>
<th>Previous week</th>
<th>Location</th>
<th>Language</th>
<th>Media-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ER</td>
<td>SR</td>
<td>ER</td>
<td>SR</td>
</tr>
<tr>
<td>Method A</td>
<td>11</td>
<td>11</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Method B</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Only Method C passes all tests successfully.
Prediction power

How well can we predict the existence of an edge given some ground truth network?

We measure AUC under a 100-folds cross validation scheme.

We measure AUC under a 100-folds cross validation scheme.
Prediction power

Using all ground truth networks together and each one separately, we get:

Method A

Method B

Method C
Prediction power

Using pairs of ground truth networks we get:

Overall, Method C gives the best results of 77.1% using all three ground truth components.
Conclusions

We showed how:

- to select the inference algorithm
- to validate the network in terms of stability, relation to ground truth components and topology prediction
Thank you!