Re-analyzing Social Network Studies
What is the enemy of my enemy?

Viviana Amati    Jürgen Lerner

Winter 2015/2016
(last updated: October 28, 2015)
Short introduction

Social network analysis and the enemy of my enemy.
What is the enemy of my enemy?

some inter-state relations in the 1980’s

Is this network pattern typical?
Social networks consist of actors and relations among them.

- **actors**: persons, organizations, companies, countries, . . .
- **relations**: friendship, asking for advice, communication, collaboration, alliances, trade, war, . . .

Actors and ties may have associated data (attributes).
Specify and estimate probability distributions for social networks.

- How likely is it that two actors are connected by a tie?
  - dependent on their attributes/behavior
  - dependent on their ties to other actors

- How likely is it that an actor has a certain attribute?
  - dependent on its ties
  - dependent on the attributes/behavior of its neighbors

Provides a framework for tackling many research questions in the social sciences.
Structural balance theory. What is the enemy of my enemy?

Structural balance theory (Heider 1946) applies to triplets of 3 actors mutually connected by positive or negative ties:

- **balanced**
- **not balanced**

SBT claims that actors prefer balanced networks.

According to structural balance theory, the enemy of my enemy should be my friend.

The topic of this seminar is to test this hypothesis in international relations networks.
Structural balance theory.
What is the enemy of my enemy?

Structural balance theory (Heider 1946) applies to triplets of 3 actors mutually connected by positive or negative ties:

- balanced
- not balanced

SBT claims that actors prefer balanced networks.

According to structural balance theory, the enemy of my enemy should be my friend.

The topic of this seminar is to test this hypothesis in international relations networks.
Outline of this seminar.

- test structural balance theory in international relations;
- test hypotheses about factors that cause imbalanced relations;
- assess the effect of imbalanced relations on the likelihood of conflict.

In this seminar we are going to re-analyze some of their research questions.
Why should we re-analyze published scientific studies.

**Fact**: Replication studies (in the social sciences) might yield different results than the original studies. This has many possible explanations . . .

While replication studies are clearly important,

- the *goal of this seminar* is to gain experience in applying statistical methods for social network analysis.

Doing a replication study is a good way to understand differences of various methods.
Why should we re-analyze published scientific studies.

**Fact**: Replication studies (in the social sciences) might yield different results than the original studies. This has many possible explanations . . .

While replication studies are clearly important,

- the **goal of this seminar** is to gain experience in applying statistical methods for social network analysis.

Doing a replication study is a good way to understand differences of various methods.
Deliverables (of this seminar).

Each participant has to provide the following deliverables:

- Implement and conduct an analysis of international relations by varying some aspects of Maoz et al. (topics will be suggested starting from next week);
- Give a presentation about your analysis and findings (talk of about 20 minutes + discussion);
- Technical report about your analysis and findings (approx. 5-10 pages; might include analysis code);
- Short summary / conclusion (approx. 2 pages) about all findings (also those of other participants) of this seminar; compare these to Maoz et al. (2007);

As well as actively participate in the seminar.
Approximate schedule (of this seminar).

**Weeks 1–5:** Short introduction to social network analysis, empirical data used in the seminar, R software for statistical computing. Analysis by participants.

**Weeks 6–11** Presentations by participants; discussion. Potentially additional short lectures.

**Weeks 12–** Spare time, wrap-up, discussion, summary.

Each participant

- choses one topic;
- has at least three weeks for doing the analysis and preparing the presentation;
- technical report is written (or at least revised) after the presentation;
- technical report and summary is due around the end of the lecture period.
First tasks until next week (October 28).

Read the paper (Maoz et al, 2007), with an emphasis on:

▶ The hypotheses RH1-RH5 on p. 102.
▶ The section "Modeling Indirect Relations" on pp. 104-105
▶ The section "Research Design" on pp. 105-108. Put most emphasis on those variables that are necessary to tackle hypotheses RH1-RH5.

Download and install R and R Studio and download the empirical data. Links are provided on http://algo.uni-konstanz.de/lehre/ws15/rsns/

IMPORTANT Before Monday (October 26): write us an email informing us if and how you want to participate.
{Viviana.Amati|Juergen.Lerner}@uni-konstanz.de
International relations networks.

Data description.
The Correlates of War project.  
http://www.correlatesofwar.org/

The Correlates of War project “seeks to facilitate the collection, dissemination, and use of accurate and reliable quantitative data in international relations.”

Provides among others longitudinal data about interstate conflicts and additional information about countries and dyads of countries.


Variables: dyad-years.

The dataset is a table. Each row encodes data for one dyad-year; each column encodes the values for one variable.

**year:** ranging from 1885 to 2001.

**dyad:** two variables (statea, stateb) give unique codes (numbers) for the two countries; (abb_a, abb_b) give 3-letter abbreviations; a pair (statea, stateb) identifies one dyad.

Triples \((t, a, b)\) for the variables year, statea, stateb are unique in the table and identify one dyad-year.

The dyads are undirected: at most one of \((t, a, b)\) or \((t, b, a)\) will be in the data (which one is meaningless).
Variables: alliances and conflicts.

**alliances:** the binary variable *allies* is 1 if *a* and *b* have at least one alliance in year *t*.

**conflicts:** the binary variable *mzmid1* is 1 if *a* and *b* fight or threaten each other in a *militarized interstate dispute* (MID) in the next(!) year *t + 1*.

These two variables are used as outcome variables and as explanatory variables. For instance, the probability of a conflict on dyad (*a, b*) might be explained by alliances and/or conflicts on dyads (*a, c*) and/or (*b, c*).
Variables: geographic closeness.

**Contiguity:** the variable `contig` encodes whether countries $a$ and $b$ are geographically adjacent. Values:
- 1 separated by a land or river border;
- 2 separated by 12 miles of water or less;
- 3 separated by 24 miles of water or less;
- 4 separated by 150 miles of water or less;
- 5 separated by 400 miles of water or less;
- 6 further apart (non-adjacent)

**Distance:** encodes the capital to capital distance.

Both variables have a strong effect on the probability to interact at all.
Variables: major powers.

Variables `majpow_a` and `majpow_b` indicate whether `a` (respectively `b`) are major powers.

<table>
<thead>
<tr>
<th>state</th>
<th>start year</th>
<th>end year</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>1898</td>
<td>2011</td>
</tr>
<tr>
<td>UKG</td>
<td>1816</td>
<td>2011</td>
</tr>
<tr>
<td>FRN</td>
<td>1816</td>
<td>1940</td>
</tr>
<tr>
<td>FRN</td>
<td>1945</td>
<td>2011</td>
</tr>
<tr>
<td>GMY</td>
<td>1816</td>
<td>1918</td>
</tr>
<tr>
<td>GMY</td>
<td>1925</td>
<td>1945</td>
</tr>
<tr>
<td>GMY</td>
<td>1991</td>
<td>2011</td>
</tr>
<tr>
<td>AUH</td>
<td>1816</td>
<td>1918</td>
</tr>
<tr>
<td>ITA</td>
<td>1860</td>
<td>1943</td>
</tr>
<tr>
<td>USR</td>
<td>1816</td>
<td>1917</td>
</tr>
<tr>
<td>USR</td>
<td>1922</td>
<td>2011</td>
</tr>
<tr>
<td>CHN</td>
<td>1950</td>
<td>2011</td>
</tr>
<tr>
<td>JPN</td>
<td>1895</td>
<td>1945</td>
</tr>
<tr>
<td>JPN</td>
<td>1991</td>
<td>2011</td>
</tr>
</tbody>
</table>

Major powers typically interact more often and also to more distant countries.
Variables: national material capabilities.

CINC: composite index of national material capabilities (derived from population size, GDP, military expenditure, steel and energy consumption).

cap_a and cap_b: CINC of state a or b.

lncaprat: logarithm of ratio of larger CINC divided by smaller CINC.

Hypothetical effect on peace and conflict: tricky.
The variable trade encodes the average of trade from $a$ to $b$ and from $b$ to $a$ (in millions current $\$). The variables opena and openb encode total trade divided by GDP of $a$ respectively $b$.

Trade relations are expected to influence the probability of alliances and conflict.
Variables: form of government.

Variables `polity_a` and `polity_b` encode the Polity II scores of country `a` (respectively `b`).

Ranges from -10 (most autocratic) to +10 (most democratic).

Two potential effects:
Democratic countries are more/less peaceful.
Countries with similar forms of governments tend to fight each other more/less.
Variables: IGO membership.

The variable \( \text{igos}_i \): interpolated number of joint intergovernmental organizations in which \( a \) and \( b \) are members.

Joint IGO membership might foster collaboration / reduce conflictive potential.
Variables: number of states.

Variable `numstate`: number of countries in the data in the given year.

Has a strong influence on the baseline probability of interaction (network *density*).
International relations networks.

Modeling alliances and conflicts.
The outcome variables: disputes and alliances.

Two countries are connected by a **negative tie** in a given year if they fight each other in a *militarized interstate dispute* (MID).

Two countries are connected by a **positive tie** in a given year if they are both members of the same formal *alliance*.

**Coding**: for two countries $u$ and $y$ and a year $t$

$$y_{uv}^{(t)} = \begin{cases} 
1 & \text{if } u \text{ and } v \text{ are allied in year } t; \\
-1 & \text{if } u \text{ and } v \text{ engage in a dyadic MID in year } t; \\
0 & \text{else.} 
\end{cases}$$

Note: ties are undirected, i.e., $y_{uv}^{(t)} = y_{vu}^{(t)}$. 
Modeling the outcome variables.

Conflicts and alliances in year $t$ are explained by data (network ties and/or covariates) from year $t - 1$.

\[
P(Y_{uv}^{(t)} = 1) = \text{functionOf(data at } t - 1)\]
\[
P(Y_{uv}^{(t)} = -1) = \text{functionOf(data at } t - 1)\]

Note: the tie between countries $u$ and $v$ is potentially also explained by data associated with other countries.

Also possible: additionally use data from years before $t - 1$.

Some advanced models explain ties in year $t$ also by other ties in year $t$ (not only by ties from the previous year).
Modeling the outcome variables by logistic regression.

Probability of a positive tie between $u$ and $v$ in year $t$:

$$P(Y_{uv}^{(t)} = 1) = \text{someFunctionOf}(\text{statistics}, \text{parameters})$$

The *statistics* (explanatory variables) quantify characteristics of the data of the previous year $t - 1$.

The *parameters* quantify the influence of those variables on the tie-probability:

- a **positive** (negative) parameter means: the higher the statistic the higher (lower) the probability;
- a **zero** parameter means: the statistic has no influence on the tie-probability.

Parameters are estimated from the empirical data.
Modeling dyadic variables by logistic regression.

Probability $p_{t;uv} = P(Y_{uv}^{(t)} = 1)$ of a positive tie from $u$ to $v$ in year $t$ specified as

$$P(Y_{uv}^{(t)} = 1) = \logit^{-1}(\theta \cdot s) = \frac{\exp(\theta \cdot s)}{\exp(\theta \cdot s) + 1},$$

where

$$s = (s_1, \ldots, s_k) \in \mathbb{R}^k \quad \text{statistics}$$

$$\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k \quad \text{parameters}$$

$$\theta \cdot s = \sum_{i=1}^{k} \theta_i \cdot s_i$$

The statistics $s_i = s_i(u, v; y^{(t-1)})$ are functions of the data at $t - 1$.

The parameters are estimated to maximize the probability of the observed sequence of networks $y = (y^{(1885)}, \ldots, y^{(2001)})$:

$$P(Y = y) = \prod_{t;uv} (p_{t;uv})^{y_{uv}^{(t)}} \cdot (1 - p_{t;uv})^{1-y_{uv}^{(t)}}.$$
Explanatory variables (I/III): ties to third actors.

Variables coding whether in year $t - 1$ there is a third actor having

- positive ties to both $u$ and $v$;
- a positive tie to $u$ and a negative tie to $v$ (or vice versa);
- negative ties to both $u$ and $v$.

These variables are related to our research questions.
Explanatory variables (II/III): inertia and degrees.

**Inertia:** variables coding whether there was a positive or negative tie between $u$ and $v$ in the preceding year $t - 1$.

**Degree:** variables coding the numbers of alliances or MIDs of $u$ and $v$ in year $t - 1$. Models differences in activity, popularity, hostility, or cooperativeness.

These are used as control variables (not directly related to our research questions).
Explanatory variables (III/III): covariates.

Variables encoding geographic closeness, distance, major power status, national material capability, trade, form of government, IGO memberships, and number of states in the system.

Also used as control variables.
Exponential random graph models (ERGM).
Allowing dependency of ties within the same year.

An ERGM specifies probabilities \( P(Y = y) \) of whole networks \( y = (y_{uv}) \), rather than individual dyadic variables.

\[
P(Y = y) = \frac{1}{z} \exp \left( \sum_{i=1}^{k} \theta_i \cdot s_i(y) \right)
\]

where

- the *statistics* \( s_i(y) \) encode numerical properties of the network \( y \);
- the *parameters* \( \theta_i \in \mathbb{R} \) model the influence of the statistics on the network probability
- \( z \) is a *normalizing constant* (making probabilities sum up to one).

Estimation is implemented in the R package **ergm**.