# Temporal Analysis of Dynamic Collaboration Graphs of Open Source Software Development: Forking

Preliminary Exam Proposal
Amir Azarbakht

# INTRODUCTION

- What: Free/Open Source Software (FOSS) development
- How: Communities, collaboration, communication
  - Artifacts, life of a FOSS project
- Community splits | Forking:
  "***When a part of a development community (or a third party not related to the project) starts a completely independent line of development based on the source code basis of the project.***"[40]

# INTRODUCTION
## Community splits | Forking

- Nature
- Frequency
- How community deals with forking, developers' opinions about it [36]
- Consequences of fork: sometime positive, sometime negative
- Sustainability (+), Dilution (-), Redundant work (-)

# PREVIOUS RESEARCH

## Retrospect: Leaders vs. Losers

- **Visual exploration** of collaboration networks of the development of WebKit [49]
- How key events in mobile-device industry affected the WebKit community over time
- Coopetition (both competition and collaboration)
- Google, Apple, Samsung played **central** roles in the WebKit network → by 2013 became leaders of mobile-device industry
- Whereas **peripheral** firms such as RIM, and Nokia → lost market-share

# PREVIOUS RESEARCH

- **Identification** [40]
  - 220 significant FOSS projects forked since 1990
  - Year, reason

| Reason for forking | Category | Abbreviation |
|---|---|---|
| Differences among developer team | Undesirable | U. F. |
| More community-driven development | Undesirable | U. F. |
| Technical differences (addition of functionality) | Other (Healthy/Other) | H. F. |
| Discontinuation of the original project | Other (not socially-related) | O. |
| Commercial strategy forks | Other (not socially-related) | O. |
| Legal issues | Other (not socially-related) | O. |

# PREVIOUS RESEARCH

## FORKING IS **EXPENSIVE**

- **Post-forking porting** of new features or bug fixes from peer projects in BSD family (FreeBSD, OpenBSD, and NetBSD) [7]
- 10-15% of lines in BSD release patches are ported edits
- 26-58% of active developers take part in porting per release
- 50% of ported changes propagate to other projects within 3 releases

Forking is expensive b/c of redundant work.

# PREVIOUS RESEARCH
# POST-HOC

The **run-up** to the forking events are seldom studied

- Was it a long-term trend?
- Was the community polarized, before forking happened?
- Was there a shift of influence? Did the center of gravity of the community change?
- What was the tipping point?
- Was it predictable? Is it ever predictable?

We are missing that context.

# PROPOSED RESEARCH CONTRIBUTION

- Study of **run-up** to forking, not post-hoc
- We treat collaboration networks as **dynamic** graphs
- Previous research gives no evidence to **identify key indicators** to predict future; we do
- **Predict forking behavior**, either good, or detrimental behavior
- Why it matters

# RESEARCH QUESTIONS

- R.Q. 1 Do forks leave traces in the collaboration artifacts of open source projects in the period leading up to the fork? (exists?)
- R.Q. 2 Do different types of forks leave different types of traces? (distinguishable?)
- R.Q. 3 What are the key indicators that let us distinguish between different types of forks? (indicators?)
- R.Q. 4 Does our analysis match what people in the community remember? (Sanity check)

# METHODOLOGY
## DATA SOURCES

- If we want to predict, we need to look at the record:
  - Developers' Mailing lists (emailed → interacted)
  - Issue(bug) tracking system (worked on same bug → interacted)
  - Source code repositories (worked on same source file → interacted)

# METHODOLOGY
## DATA COLLECTION

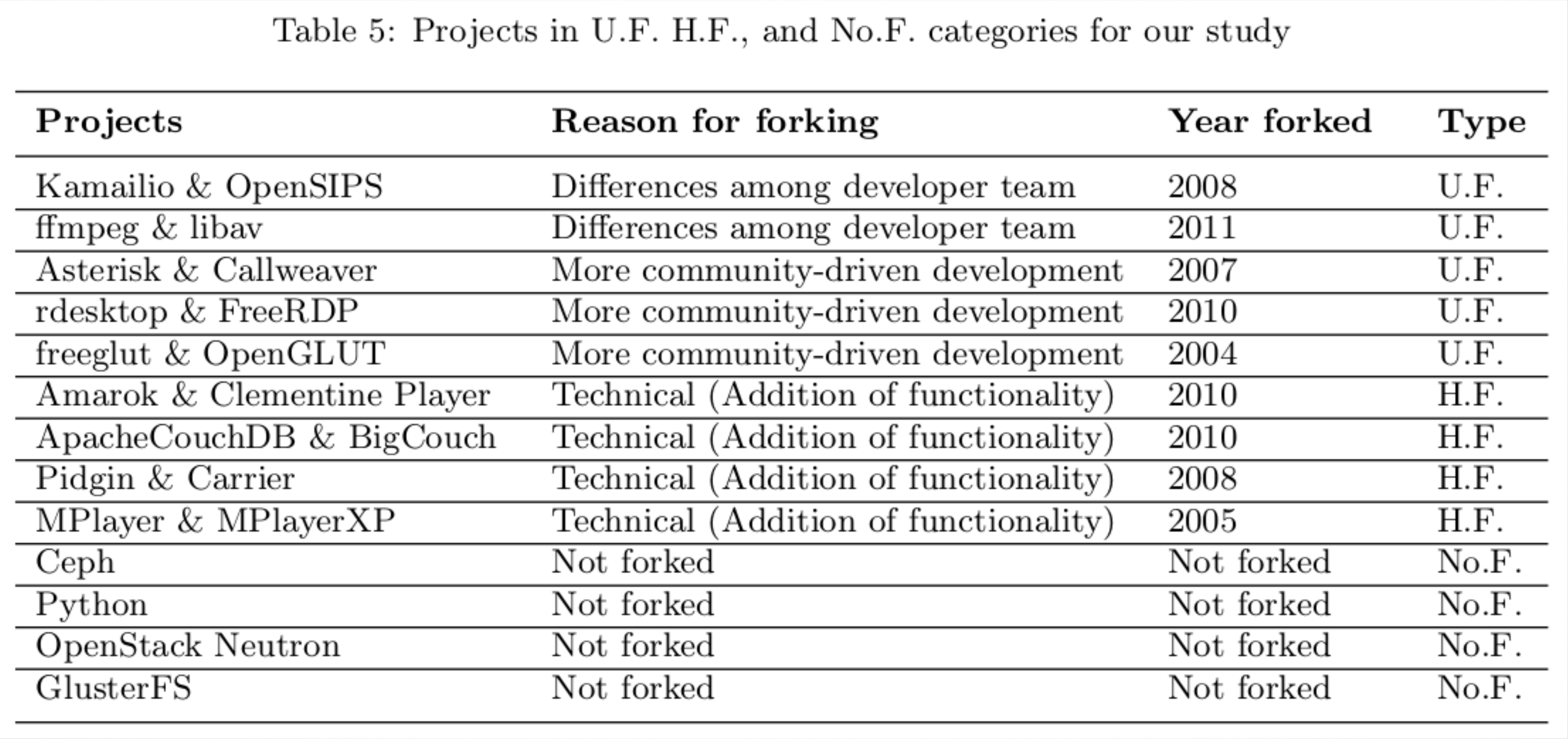


Table 5: Projects in U.F. H.F., and No.F. categories for our study

| Projects | Reason for forking | Year forked | Type |
|---|---|---|---|
| Kamailio & OpenSIPS | Differences among developer team | 2008 | U.F. |
| ffmpeg & libav | Differences among developer team | 2011 | U.F. |
| Asterisk & Callweaver | More community-driven development | 2007 | U.F. |
| rdesktop & FreeRDP | More community-driven development | 2010 | U.F. |
| freeglut & OpenGLUT | More community-driven development | 2004 | U.F. |
| Amarok & Clementine Player | Technical (Addition of functionality) | 2010 | H.F. |
| ApacheCouchDB & BigCouch | Technical (Addition of functionality) | 2010 | H.F. |
| Pidgin & Carrier | Technical (Addition of functionality) | 2008 | H.F. |
| MPlayer & MPlayerXP | Technical (Addition of functionality) | 2005 | H.F. |
| Ceph | Not forked | Not forked | No.F. |
| Python | Not forked | Not forked | No.F. |
| OpenStack Neutron | Not forked | Not forked | No.F. |
| GlusterFS | Not forked | Not forked | No.F. |

# METHODOLOGY
## GRAPH REPRESENTATION

- Representation:
  - Static (snapshot) vs. **dynamic** sociograms

- Analysis:
  - Cross-sectional vs. **Longitudinal**

# PREDICTIONS

- To help form predictions, we turn to social sciences as bases for our hypotheses. Three important theories:
  - **Balance Theory**
  - **Signaling Theory**
  - **Assortativity Theory**

# PREDICTIONS THEORIES

- **Balance Theory**
  - Motivation of individuals to move toward psychological balance
  - Cognitive consistency drives sentiment or liking relationships, & liking of things created by or associated with the alter in the relationship
  - Prediction: in U.F. personal, cognitive inconsistency between liking relationship of other developers & the software and community created by or associated with them → subcommunity leaves

# PREDICTIONS THEORIES

- **Signaling Theory**
  - Communication between individuals in terms of signals
  - Helps the needy only if helping signals desirable personality. Cost of helping should be less than benefit gained b/c of signal
  - Prediction: U.F. more community-driven, decrease in preferential attachment, because of lowering tendency to help others and gain social prestige

# PREDICTIONS
# THEORIES

- **Assortativity Theory**
  - Collaboration,
  - People’s preference for interacting with others who are similar to them in some way
    - Homophily vs. Heterophily
  - Related to reciprocity (contingent, direct, indirect)
  - Prediction: U.F. more community-driven, increase in assortativity as a by-result of homophily.

# MODEL COVARIATES & EXPECTATIONS

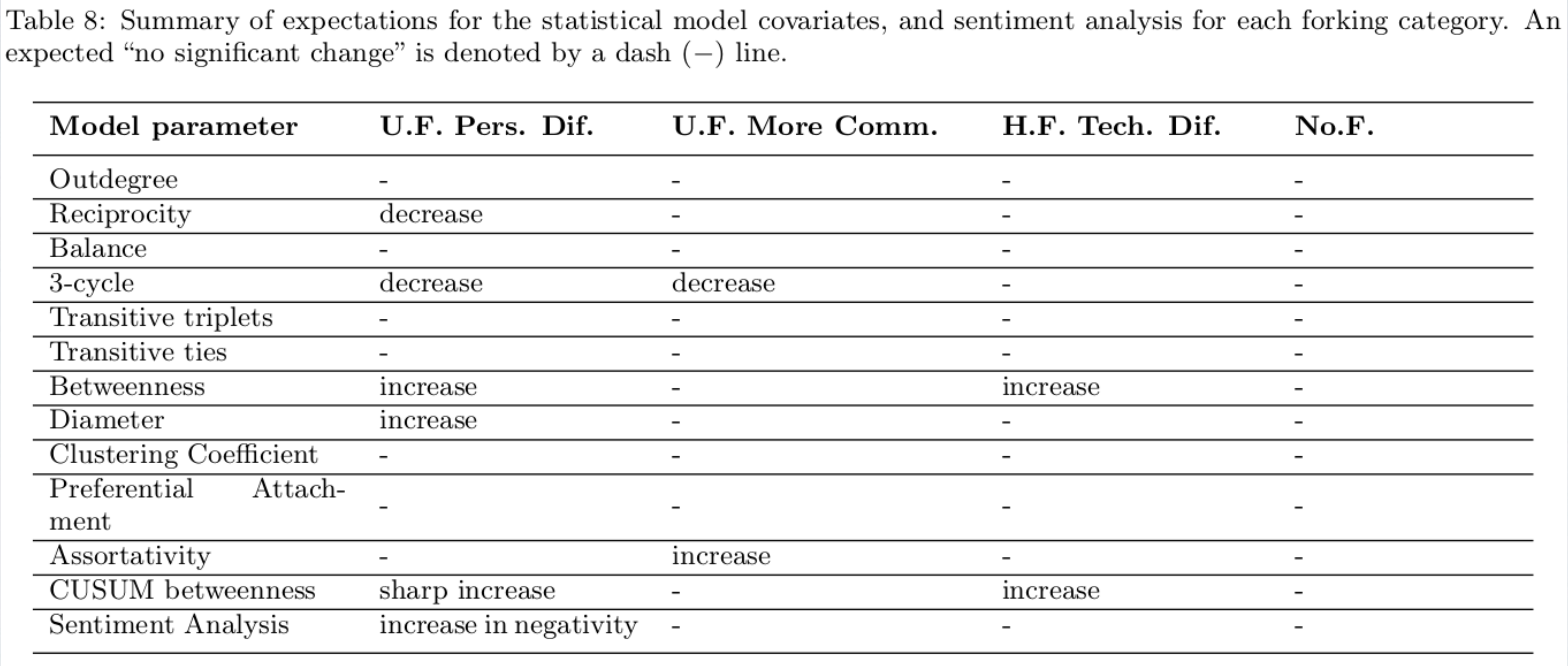


Table 8: Summary of expectations for the statistical model covariates, and sentiment analysis for each forking category. An expected "no significant change" is denoted by a dash (−) line.

| Model parameter | U.F. Pers. Dif. | U.F. More Comm. | H.F. Tech. Dif. | No.F. |
|---|---|---|---|---|
| Outdegree | - | - | - | - |
| Reciprocity | decrease | - | - | - |
| Balance | - | - | - | - |
| 3-cycle | decrease | decrease | - | - |
| Transitive triplets | - | - | - | - |
| Transitive ties | - | - | - | - |
| Betweenness | increase | - | increase | - |
| Diameter | increase | - | - | - |
| Clustering Coefficient | - | - | - | - |
| Preferential Attachment | - | - | - | - |
| Assortativity | - | increase | - | - |
| CUSUM betweenness | sharp increase | - | increase | - |
| Sentiment Analysis | increase in negativity | - | - | - |

- Why

# YET TO BE DONE

We proposed to model the developers interactions statistically, to find the population processes that underlie the formation, changes, and dissolution of developer communities. We expect to see distinct changes of such processes for each forking category.

- Data collection for mailing list archives is completed.
- The issue tracking system, and source code interactions data is in the progress.
- Once all data is collected an cleaned, we will do the statistical modeling.
- Next, we will conduct the interview study, to answer R.Q. 4.
- We expect to find patterns specific to each category, which then may be used to identify early warning signs of forking. The identification of such measures may inform those who are interested in the sustainability of their project community to stay informed and take action to amend undesirable dynamics.

# Timeline

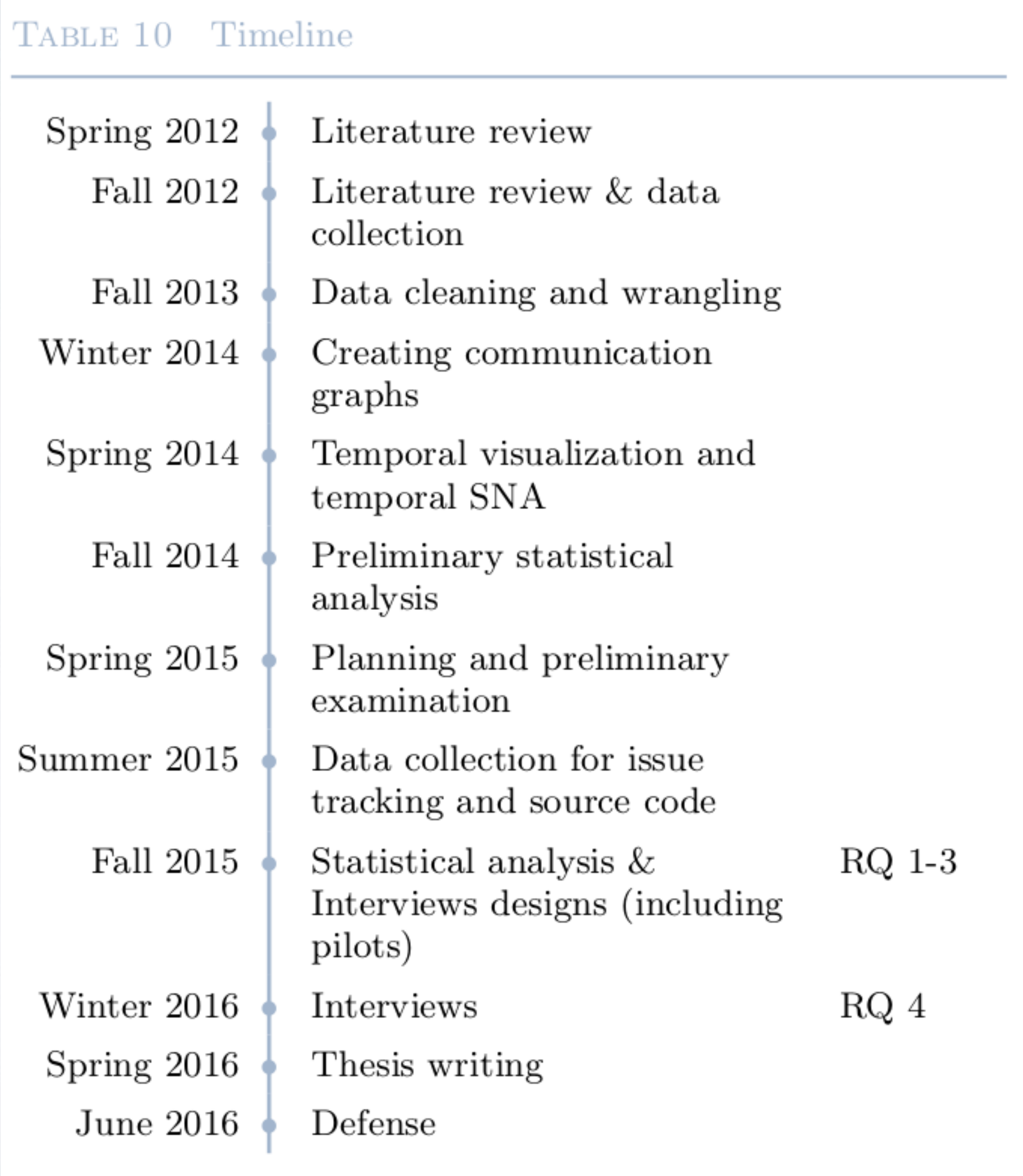


TABLE 10 Timeline

| | | |
|---|---|---|
| Spring 2012 | Literature review | |
| Fall 2012 | Literature review & data collection | |
| Fall 2013 | Data cleaning and wrangling | |
| Winter 2014 | Creating communication graphs | |
| Spring 2014 | Temporal visualization and temporal SNA | |
| Fall 2014 | Preliminary statistical analysis | |
| Spring 2015 | Planning and preliminary examination | |
| Summer 2015 | Data collection for issue tracking and source code | |
| Fall 2015 | Statistical analysis & Interviews designs (including pilots) | RQ 1-3 |
| Winter 2016 | Interviews | RQ 4 |
| Spring 2016 | Thesis writing | |
| June 2016 | Defense | |

# End of Presentation

# THANK YOU.

Questions?

# METHODOLOGY

## Phase 2: Statistical Modeling

- Observed network vs. Population it belongs to
- Analysis approaches:
  - Network-specific vs. Population processes

# METHODOLOGY

## Why bother finding a statistical model?

- Observed network: observation error, uncertainty -> perturbation in numeric descriptors

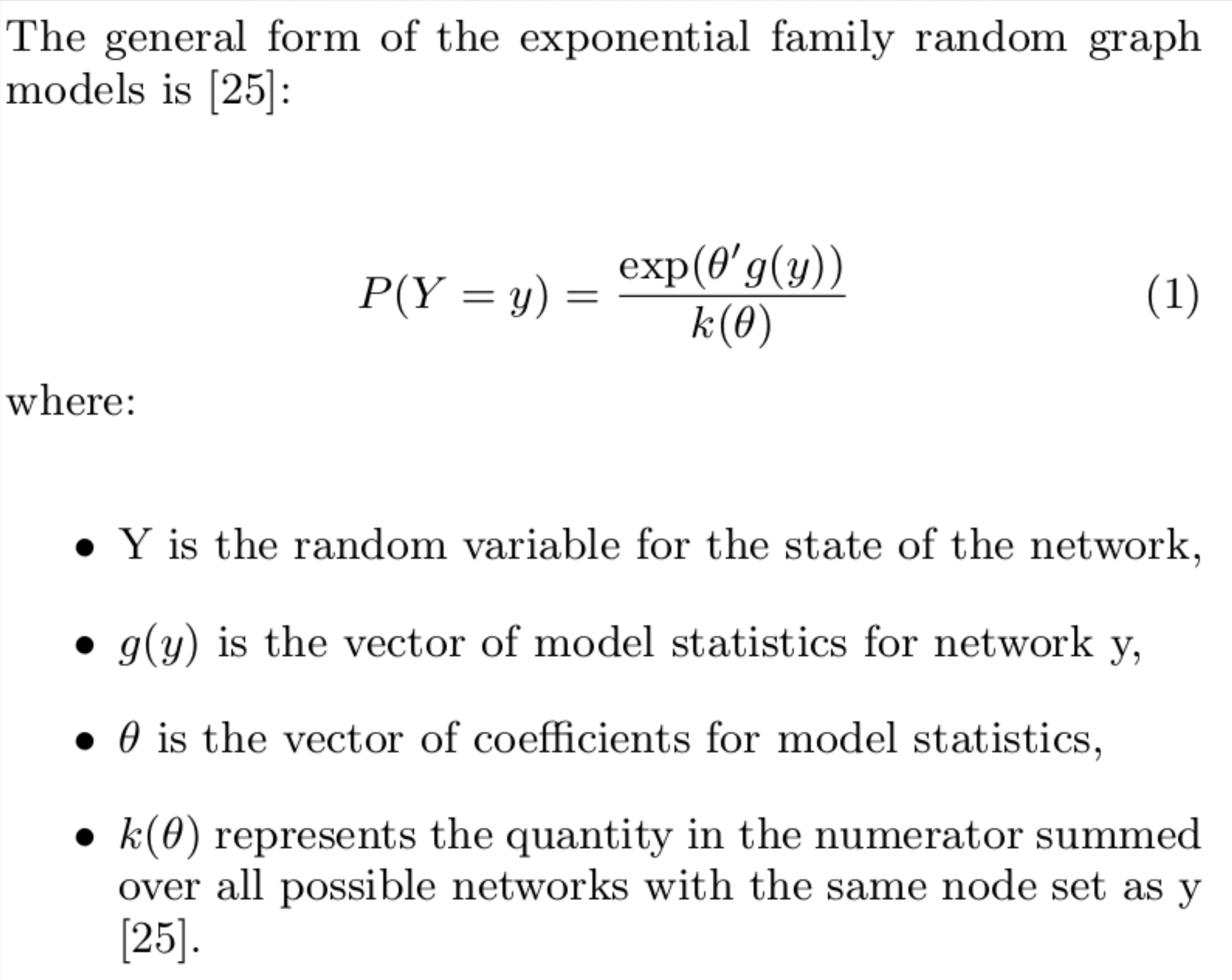
- Network-specific assumes edge independence; social data is relational (Balance Theory)

- Population-specific identifies social forces that helped form the network; simulation; finds statistical distribution of population; compare; significance vs. noise

- Different social processes may manifest similar network structures. Structural vs. Node-level

# THE STATISTICAL MODEL

## Exponential family random graph models (ERGM)



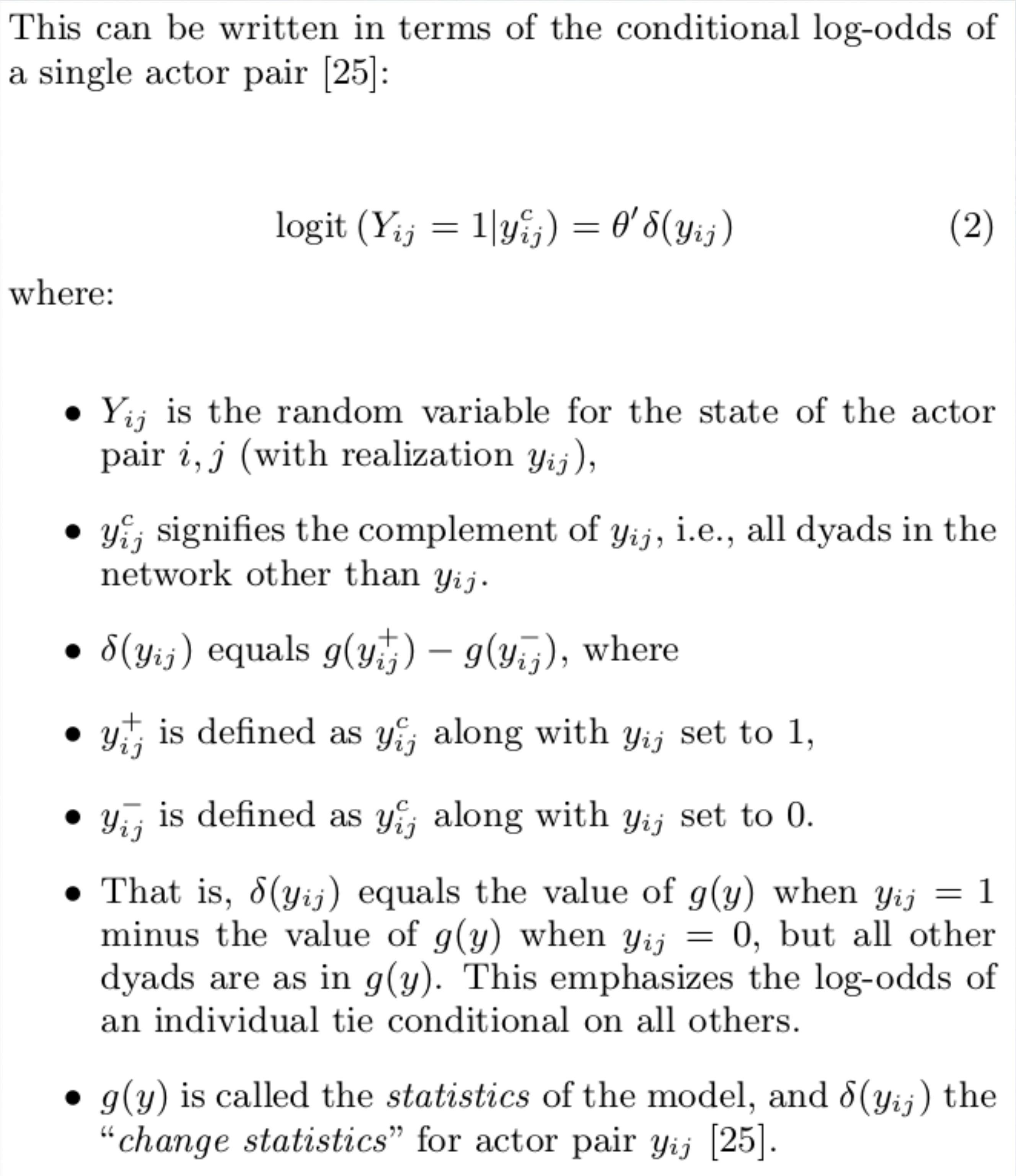
The general form of the exponential family random graph models is [25]:

$$P(Y = y) = \frac{\exp(\theta' g(y))}{k(\theta)} \tag{1}$$

where:

- Y is the random variable for the state of the network,
- $g(y)$ is the vector of model statistics for network y,
- $\theta$ is the vector of coefficients for model statistics,
- $k(\theta)$ represents the quantity in the numerator summed over all possible networks with the same node set as y [25].

# THE STATISTICAL MODEL



This can be written in terms of the conditional log-odds of a single actor pair [25]:

$$\text{logit}\,(Y_{ij} = 1 | y_{ij}^{c}) = \theta' \delta(y_{ij}) \tag{2}$$

where:

- $Y_{ij}$ is the random variable for the state of the actor pair $i, j$ (with realization $y_{ij}$),
- $y_{ij}^{c}$ signifies the complement of $y_{ij}$, i.e., all dyads in the network other than $y_{ij}$.
- $\delta(y_{ij})$ equals $g(y_{ij}^{+}) - g(y_{ij}^{-})$, where
- $y_{ij}^{+}$ is defined as $y_{ij}^{c}$ along with $y_{ij}$ set to 1,
- $y_{ij}^{-}$ is defined as $y_{ij}^{c}$ along with $y_{ij}$ set to 0.
- That is, $\delta(y_{ij})$ equals the value of $g(y)$ when $y_{ij} = 1$ minus the value of $g(y)$ when $y_{ij} = 0$, but all other dyads are as in $g(y)$. This emphasizes the log-odds of an individual tie conditional on all others.
- $g(y)$ is called the *statistics* of the model, and $\delta(y_{ij})$ the "*change statistics*" for actor pair $y_{ij}$ [25].

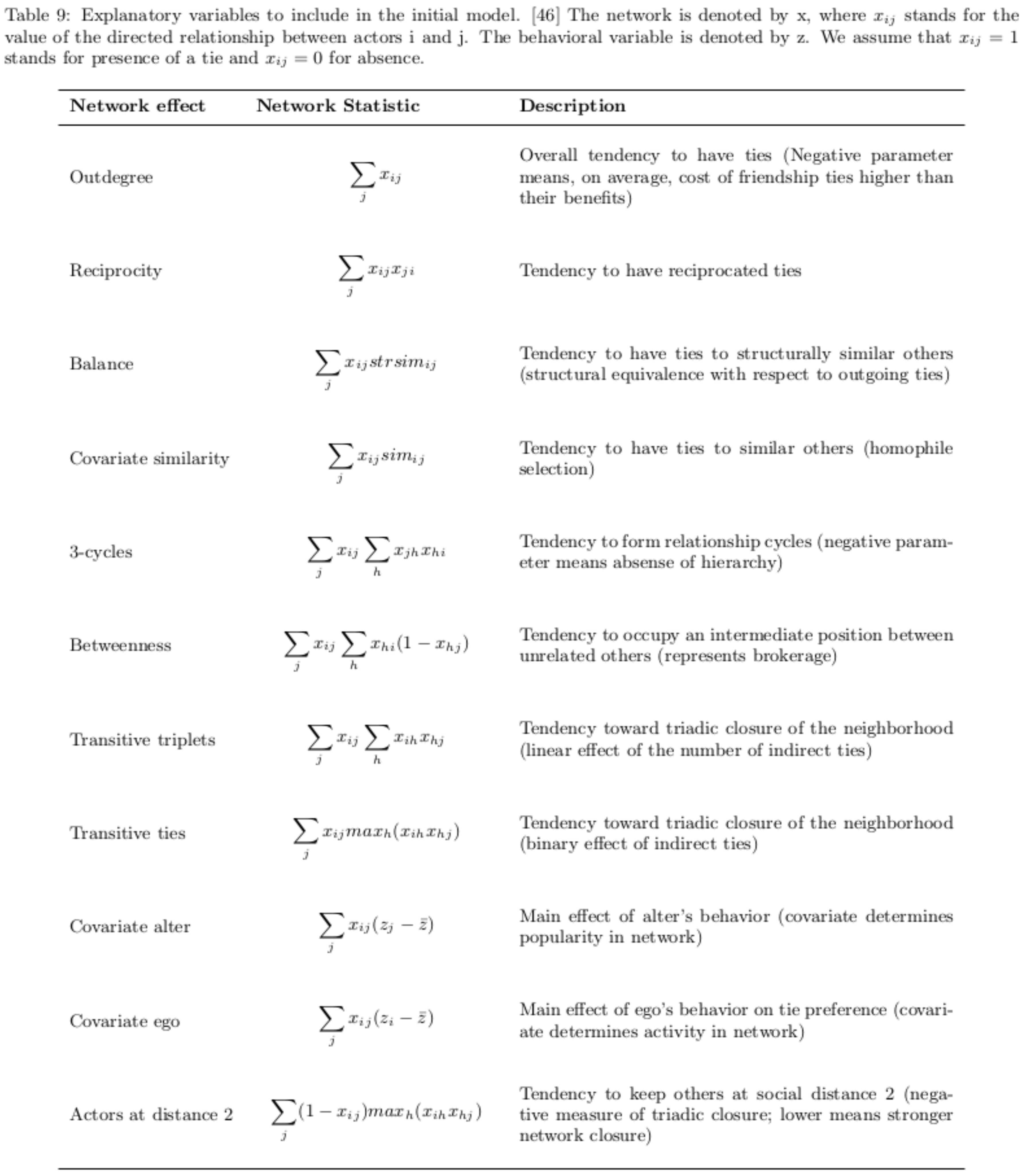


Table 9: Explanatory variables to include in the initial model. [46] The network is denoted by x, where $x_{ij}$ stands for the value of the directed relationship between actors i and j. The behavioral variable is denoted by z. We assume that $x_{ij} = 1$ stands for presence of a tie and $x_{ij} = 0$ for absence.

| Network effect | Network Statistic | Description |
|---|---|---|
| Outdegree | $\sum_j x_{ij}$ | Overall tendency to have ties (Negative parameter means, on average, cost of friendship ties higher than their benefits) |
| Reciprocity | $\sum_j x_{ij}x_{ji}$ | Tendency to have reciprocated ties |
| Balance | $\sum_j x_{ij} strsim_{ij}$ | Tendency to have ties to structurally similar others (structural equivalence with respect to outgoing ties) |
| Covariate similarity | $\sum_j x_{ij} sim_{ij}$ | Tendency to have ties to similar others (homophile selection) |
| 3-cycles | $\sum_j x_{ij} \sum_h x_{jh}x_{hi}$ | Tendency to form relationship cycles (negative parameter means absense of hierarchy) |
| Betweenness | $\sum_j x_{ij} \sum_h x_{hi}(1 - x_{hj})$ | Tendency to occupy an intermediate position between unrelated others (represents brokerage) |
| Transitive triplets | $\sum_j x_{ij} \sum_h x_{ih}x_{hj}$ | Tendency toward triadic closure of the neighborhood (linear effect of the number of indirect ties) |
| Transitive ties | $\sum_j x_{ij} max_h(x_{ih}x_{hj})$ | Tendency toward triadic closure of the neighborhood (binary effect of indirect ties) |
| Covariate alter | $\sum_j x_{ij}(z_j - \bar{z})$ | Main effect of alter's behavior (covariate determines popularity in network) |
| Covariate ego | $\sum_j x_{ij}(z_i - \bar{z})$ | Main effect of ego's behavior on tie preference (covariate determines activity in network) |
| Actors at distance 2 | $\sum_j (1 - x_{ij}) max_h(x_{ih}x_{hj})$ | Tendency to keep others at social distance 2 (negative measure of triadic closure; lower means stronger network closure) |

# Model Covariates

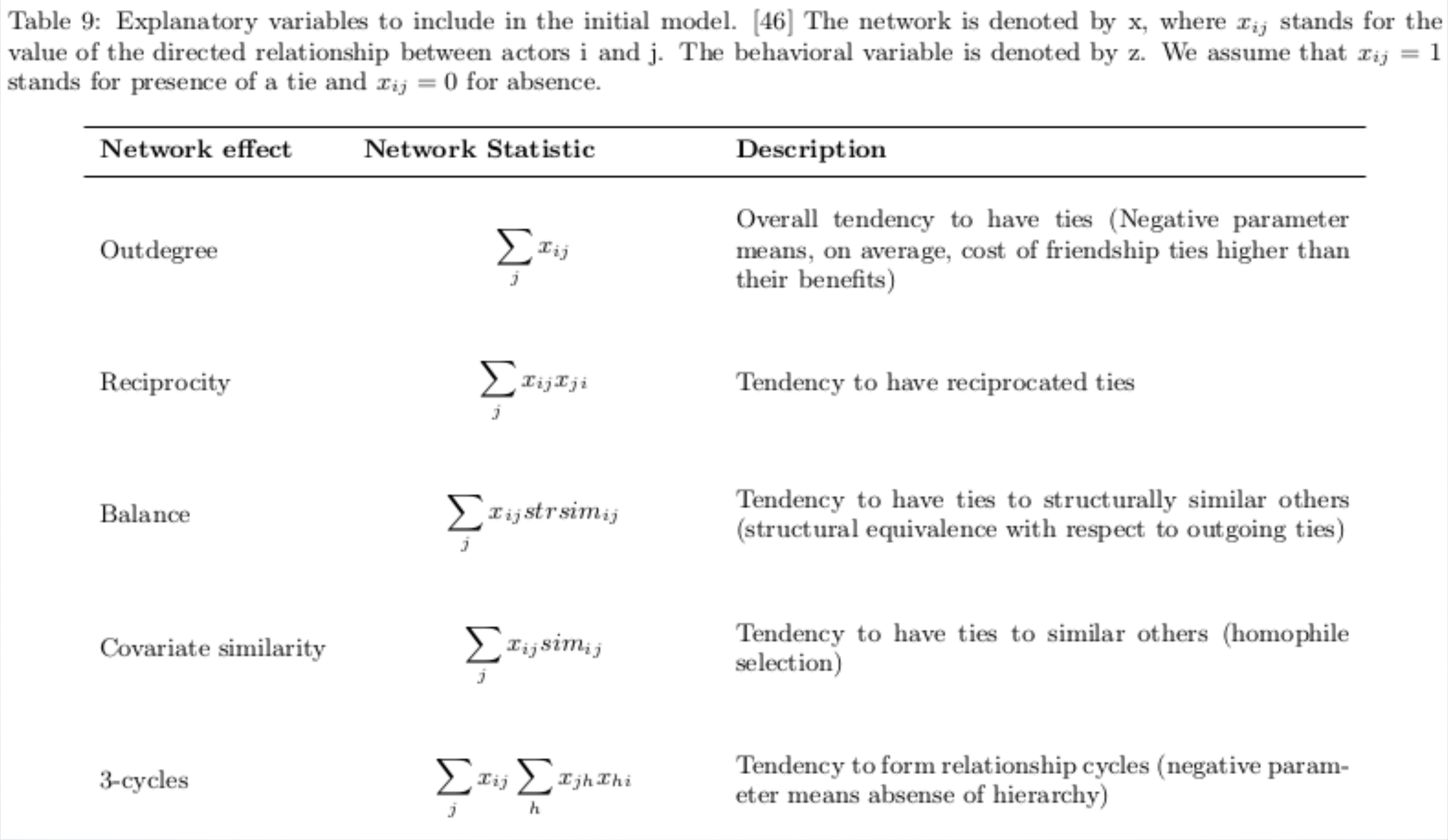
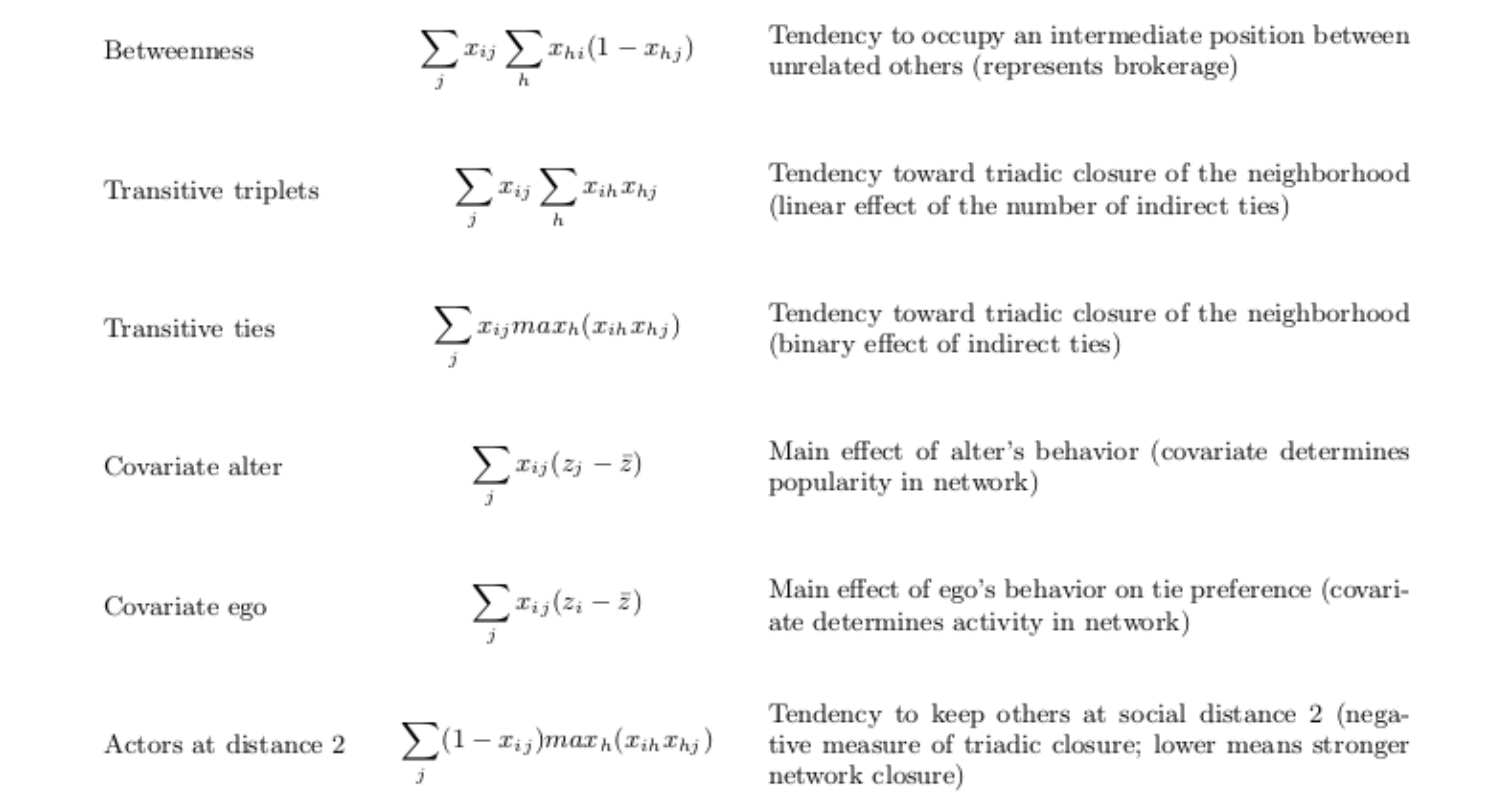


Table 9: Explanatory variables to include in the initial model. [46] The network is denoted by x, where $x_{ij}$ stands for the value of the directed relationship between actors i and j. The behavioral variable is denoted by z. We assume that $x_{ij} = 1$ stands for presence of a tie and $x_{ij} = 0$ for absence.

| Network effect | Network Statistic | Description |
|---|---|---|
| Outdegree | $\sum_j x_{ij}$ | Overall tendency to have ties (Negative parameter means, on average, cost of friendship ties higher than their benefits) |
| Reciprocity | $\sum_j x_{ij}x_{ji}$ | Tendency to have reciprocated ties |
| Balance | $\sum_j x_{ij} strsim_{ij}$ | Tendency to have ties to structurally similar others (structural equivalence with respect to outgoing ties) |
| Covariate similarity | $\sum_j x_{ij} sim_{ij}$ | Tendency to have ties to similar others (homophile selection) |
| 3-cycles | $\sum_j x_{ij} \sum_h x_{jh}x_{hi}$ | Tendency to form relationship cycles (negative parameter means absense of hierarchy) |

# Model Covariates



| | | |
|---|---|---|
| Betweenness | $\sum_j x_{ij} \sum_h x_{hi}(1 - x_{hj})$ | Tendency to occupy an intermediate position between unrelated others (represents brokerage) |
| Transitive triplets | $\sum_j x_{ij} \sum_h x_{ih} x_{hj}$ | Tendency toward triadic closure of the neighborhood (linear effect of the number of indirect ties) |
| Transitive ties | $\sum_j x_{ij} max_h(x_{ih} x_{hj})$ | Tendency toward triadic closure of the neighborhood (binary effect of indirect ties) |
| Covariate alter | $\sum_j x_{ij}(z_j - \bar{z})$ | Main effect of alter's behavior (covariate determines popularity in network) |
| Covariate ego | $\sum_j x_{ij}(z_i - \bar{z})$ | Main effect of ego's behavior on tie preference (covariate determines activity in network) |
| Actors at distance 2 | $\sum_j (1 - x_{ij}) max_h(x_{ih} x_{hj})$ | Tendency to keep others at social distance 2 (negative measure of triadic closure; lower means stronger network closure) |