Models of Friendship Formation for the Generation of Synthetic Social Networks

Thibaut Dubernet, Kay W. Axhausen

Institute for Transport Planning and Systems (IVT)
ETH Zurich

DCM Workshop — EPFL, Lausanne — April 2016

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

What this Presentation is

- Presentation of Work In Progress
- Development of models of friendship formation
- Aim: *generating* social networks
- Based on previous work by Matthias Kowald and Theo Arentze
- Criticism Welcome!
  - Probability that I missed some important detail significantly different from 0

Introduction

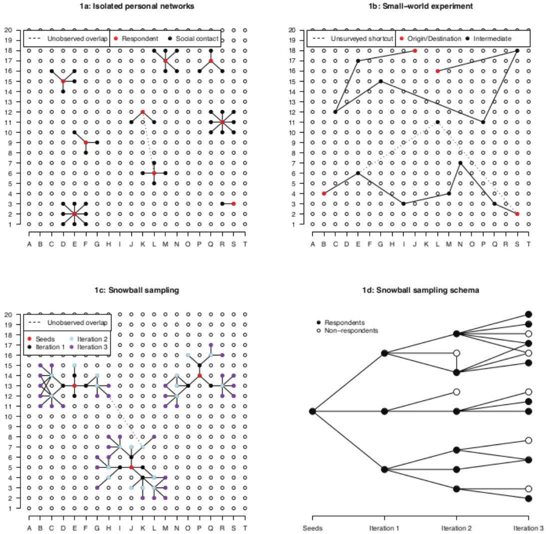
- Social contacts and their distribution assumed to have important impact on where leisure activities are performed
- If it is the case, social network data might help in forecasting
- Important characteristic of social networks:
    - Spatial distribution of social contacts
    - Homophily
    - Degree distribution
    - Transitivity/Clustering
- Generating synthetic social networks:
    - Reproduce important characteristics
    - Be computationally *scalable*
        - aim: generate network for synthetic Swiss population

A Word on the Data

- Vocabulary:
  - ego: person of interest
  - alter: (potential) friend of an ego
  - tie: existence of a relationship of interest
  - social network: graph where nodes are egos and edges are ties
  - ego-centric network: graph composed by one ego and its alters
  - clustering: proportion of possible triangle that are closed
    - "friends of friends that are friends"
- Assume we have a way to reveal (sub) network
- Static view
- In our case: snowball sample
  - Focus on *leisure contacts*
  - assumption: all relevant ties are reported (not in the dataset $\Rightarrow$ not in the real world)

Snowball Sampling

Source: Kowald (2013)



Inspiration: T. Arentze *et Al.* 2013

- Estimated on the Zurich Snowball Data
- Designed for the same purpose as here
- Only tested for a small synthetic population
- Requires calibration

Idea

- Associate a random utility to each potential tie
- The probability for a friendship to exist is the probability that this utility is higher than a fixed threshold

$$P(ij) = P(U_{ij} + \varepsilon_{ij} > u_0)$$

- Threshold is lower in case of common friends

$$P(ij) = P(U_{ij} + \varepsilon_{ij} > u_0 - \Theta)$$

  - transitivity
- U_{ij} symmetric, and contains distance and homophily measures

Pros and Cons

- Pros
  - ε_{ij} logistically distributed leads to closed form likelihood
    - basically a "yes/no" logit for each tie
  - each tie can be considered in (almost) isolation for estimation
  - intuitive two-rounds generation algorithm
- Cons
  - Degree increases with size of the "choice set"
    - thresholds u_0 and Θ need to be calibrated to reproduce average degree and clustering
  - Diversity of chosen friends decreases with size of the choice set
    - in particular spatial distribution!

Results

			Homophily		
Soc. Net.	Clust.	Avg. Deg.	Age	Gender	Dist.
Snowball	0.206	22	46.3%	61.7%	26.6 km
0.025%	0.190	22	30.7%	56.5%	49.1 km
0.025% (ZH)	0.187	20.6	29.4%	55.7%	17.8 km
10%	0.150	21.7	45.2%	66.0%	18.8 km
10% (ZH)	0.225	20.6	45.4%	66.2%	7.3 km

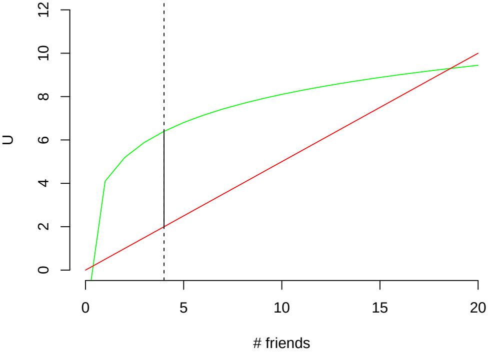
New Model

- Aim: try to overcome the Cons
  - (leads to dropping some Pros...)
- Basic Idea:
  - friends come with a utility (they are nice) and a cost (but they cost time)
  - "marginal utility" of an additional friend decreases with number of friends
  - individuals balance utility and cost
  - possible cost functions:
    - linear in ego-centric network size
    - linear in number of *cliques*
  - "multiple discreteness" formulation

Natural Formulation

- Basic decision rule: ego e choses the ego-centric network $\mathcal{N}$ that maximizes

$$\log\left(\sum_{i\in\mathcal{N}} U_{ei}\right) - \mathcal{C}(\mathcal{N})$$



Problems

- In this general form, combinatorial
- might be possible to make estimable/simulable by additional hypotheses. . .

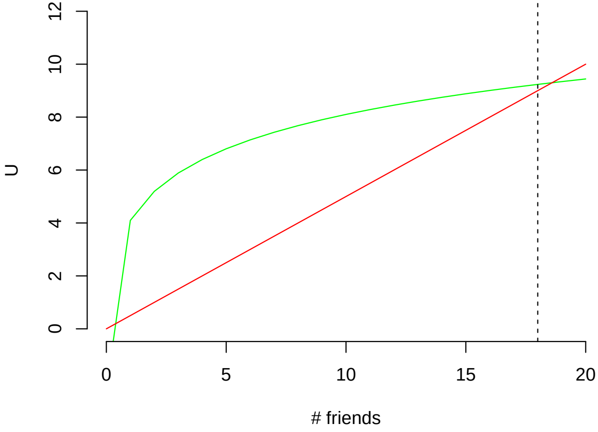
Alternative Formulation

- Basic decision rule: ego e accepts any tie as long as resulting ego-centric network $\mathcal{N}$ satisfies

$$\log\left(\sum_{i\in\mathcal{N}} U_{ei}\right) \geq \mathcal{C}(\mathcal{N})$$

- for estimation: likelihood of a tie is its probability of acceptance *given the rest of the ego-centric network*
- for simulation:
  - simulate the utilities
  - greedy algorithm: grow the network until no improvement is possible
    - select all remaining agents in the "choice set" that fulfill de condition
    - take one at random
    - should work if $\mathcal{C}(\mathcal{N})$ grows with $|\mathcal{N}|$

Intuition



Likelihood: Probability of a Tie

- Consider each tie independently
  - non-realized ties as well
- $U_{ea} = U_{ae} = V_{ea} + \varepsilon_{ea}$
- ego e accepts tie ea if

$$\varepsilon_{ea} > \exp\left(\mathcal{C}(\mathcal{N}^{e}_{+ea})\right) - \sum_{i \neq a} U_{ei} - V_{ea} = \Theta_{ea}$$

- we know the existing network, so we know

$$\sum_{i \neq a} U_{ei} > \exp(\mathcal{C}(\mathcal{N}^{e}_{-ea})) = \Theta_{-ea}$$

- The probability $P(ea)$ to observe a tie ea is:

$$P\left(\varepsilon_{ea} > \max(\Theta_{ea}, \Theta_{ae}) \middle| \sum_{i \neq a} U_{ei} > \Theta_{-ea}, \sum_{i \neq e} U_{ai} > \Theta_{-ae}\right)$$

Estimation

- No chance of getting a closed form here
- Assume $\varepsilon \sim \mathcal{N}(0, 1)$. Then the sum of n realizations follows $\mathcal{N}(0, n)$
- For each tie, can simulate $P(ea)$ (resp. $P(\neg ea)$):
  - Sampling $\sum_{i \neq a} U_{ei}$ from truncated normal distribution
  - Inject resulting Θ_{ea} in $1 - \mathrm{CDF}$
  - Average
  - Likelihood: $\Pi_{ea \in obs} P(ea) \Pi_{ea \in \neg obs} P(\neg ea)$
- First results
  - use `R` and `maxLik` package
  - First estimation still running...
  - Quite expensive computationally
  - Generation: `Java` code from the Arentze approach largely usable

Pros and Cons

- Pros
  - each tie can be considered in (almost) isolation for estimation
  - intuitive generation algorithm?
  - no calibration?
  - degree should be relatively stable with choice set size
  - diversity of generated social networks should be relatively stable
- Cons
  - no closed form likelihood
  - others to discover...

Conclusions

- Design of a Model to generate social networks
- Existing model works, but has important flaws
- Tradeoff between elegance and usability
- Actual interest of the model still to be tested...