Paper: bit.ly/www19-c2g

Choosing to Grow a Graph

Modeling Network Formation as Discrete Choice

Jan Overgoor
Stanford University
overgoor@stanford.edu

Austin Benson
Cornell University
arb@cs.cornell.edu

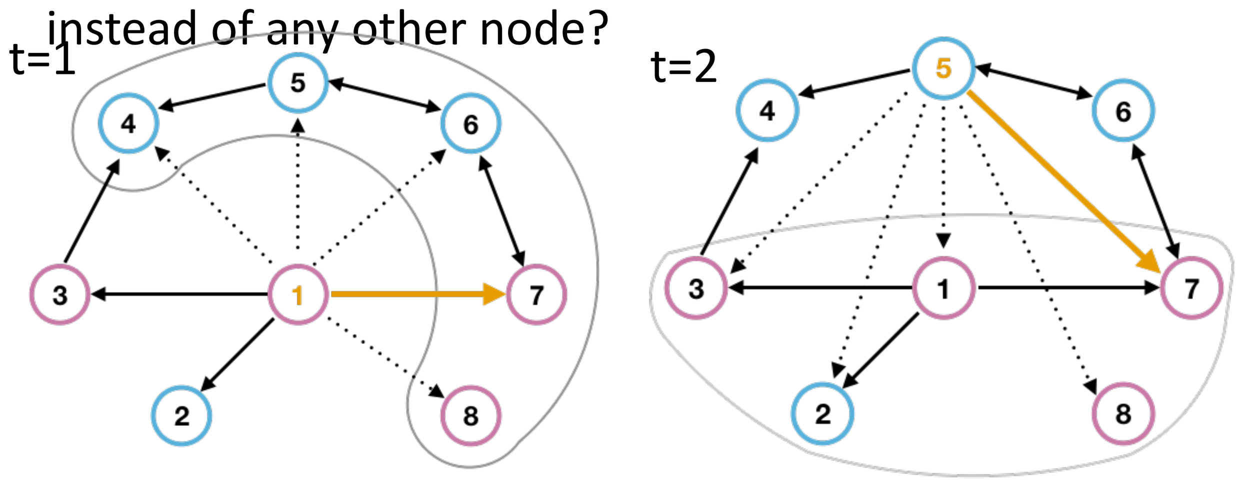
Johan Ugander
Stanford University
jugander@stanford.edu

Introduction

- Network evolution is widely studied and many different models and frameworks have been proposed.
- We frame edge formation as a **discrete choice process**, subsuming many existing models in a unified framework.
- This perspective is general, flexible, easily extended and efficiently estimated with existing analysis tools.

Discrete Choice

- Choice models are commonly used in other fields to model how individuals make choices from a slate of discrete alternatives. Alternatives have features and choice models aim to estimate the relative importance of such features.
- Edge formation events in social networks can be viewed as discrete choices. Why did i choose to connect to j instead of any other node?



- We focus on the **conditional logit model**:

$$P_i(j,C) = \frac{\exp u_{i,j}}{\sum_{\ell \in C} \exp u_{i,\ell}} = \frac{\exp \theta^T x_j}{\sum_{\ell \in C} \exp \theta^T x_\ell}$$

- The logit is a random utility model (RUM), s.t. choices are interpretable as a rational actor acting based on the "utility" sampled from random variables that decompose into the inherent utility of the alternative and a noise term.
- We can use existing optimization routines to estimate model parameters and existing statistical methods to asses the uncertainty of the estimates.

Models

- Here are a number of prior proposed models for network growth, and their corresponding functional forms to estimate each one using the conditional logit

Process	$u_{i,j}$	C
Uniform attachment	1	V
Preferential attachment	$\alpha \log d_j$	V
Non-parametric PA	θ_{d_j}	V
Triadic closure	1	$\{j : FoF_{i,j}\}$
FoF attachment	$\alpha \log \eta_{i,j}$	V
PA, FoFs only	$\alpha \log d_j$	$\{j : FoF_{i,j}\}$
Individual node fitness	θ_j	V
PA with fitness	$\alpha \log d_j + \theta_j$	V
Latent space	$\beta \cdot d(i,j)$	V
Stochastic block model	ω_{g_i, g_j}	V
Homophily	$h \cdot \mathbb{1}\{g_i = g_j\}$	V

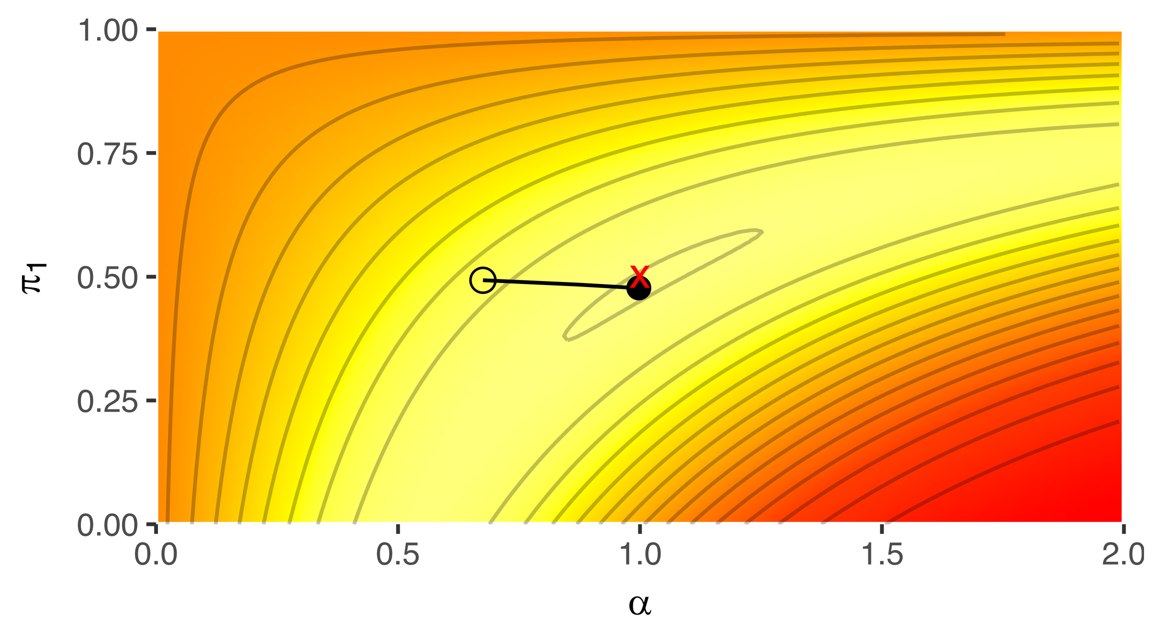
Data

- We assume that we have access to a sequence of directed edges (*i, j, t*), in chronological order.
- For every edge, create a data point for every alternative with their features at time t and whether they got selected

Choice ID	i	j	Color	$\deg_{j,t}$	$FoF_{ij,t}$	Y
1	1	4	●	2	1	0
1	1	5	●	1	0	0
1	1	6	●	2	0	0
1	1	7	●	1	0	1
1	1	8	●	0	0	0
2	5	1	●	0	0	0
2	5	2	●	1	0	0
2	5	3	●	1	0	0
2	5	7	●	2	1	1
2	5	8	●	0	0	0

Estimation

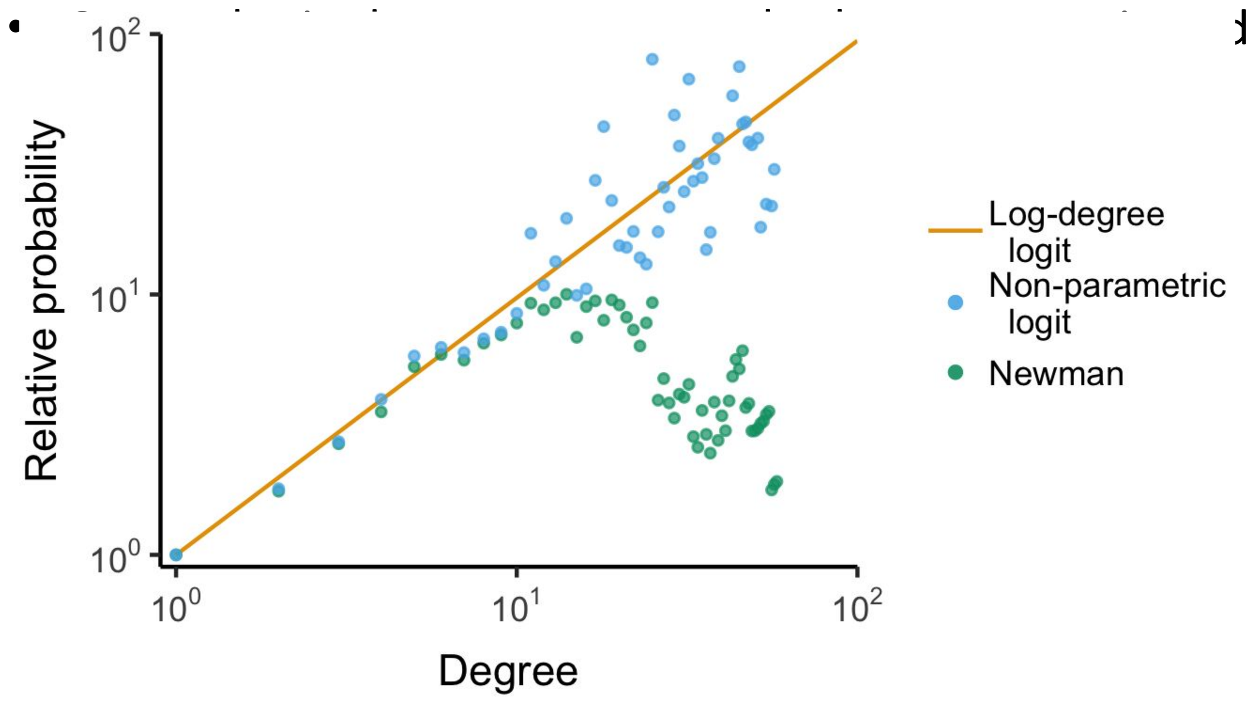
- Logit models with linear utility have a convex (wrt. θ) likelihood function and **can be efficiently maximized using standard gradient-based optimization** (e.g., BFGS). The functional form of the logit has simple gradients.



- There are a number of existing software packages (e.g. **mlogit**, **statsmodels**) to fit these models as well.
- For large sparse graphs, the choice sets can become prohibitively large. A reduced data set can be created by **sampling *s* negative/non-chosen examples**.
- When negative samples are sampled uniformly at random, parameter estimates on the sampled data are **unbiased and consistent** for the estimates on the on the full set.

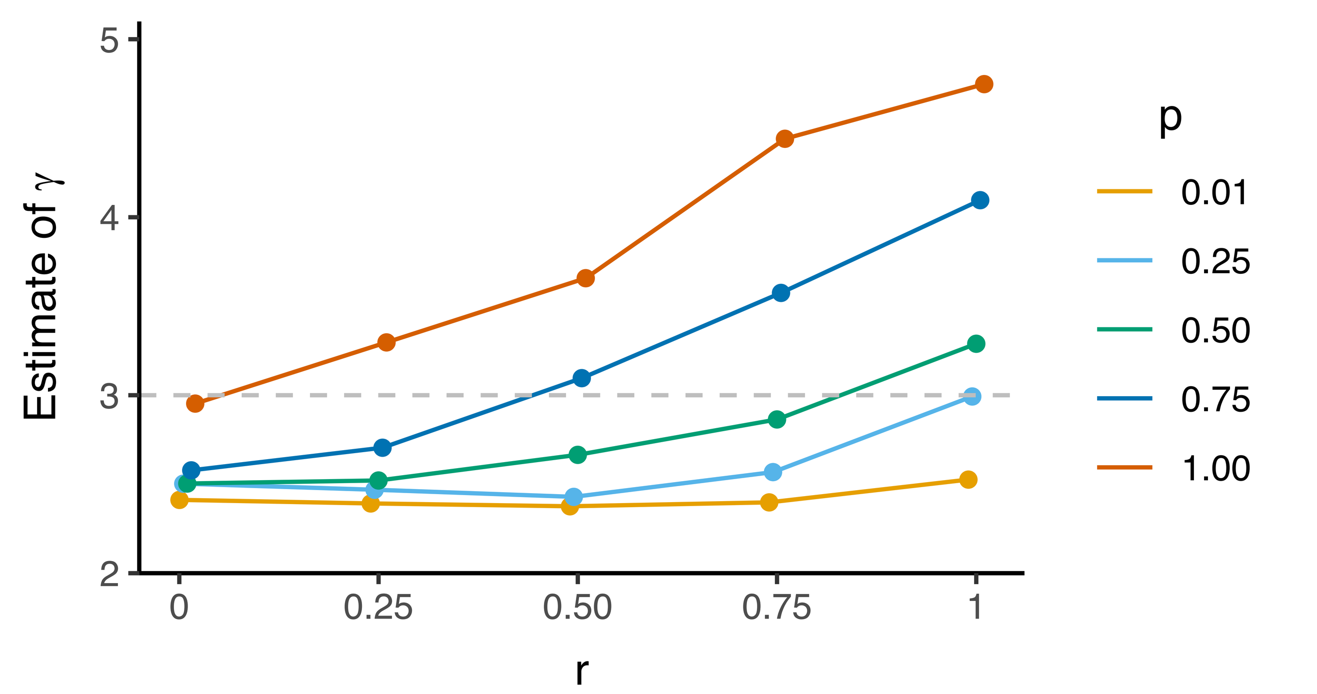
Application – Measuring PA

- The conditional logit framework provides a principled and flexible statistical test for the presence of hypothesized tendencies in a network formation process.
- For example, the presence of preferential attachment (PA), is tested when the utility specification includes $\alpha \log d_j$.



Application – PA vs Triadic Closure

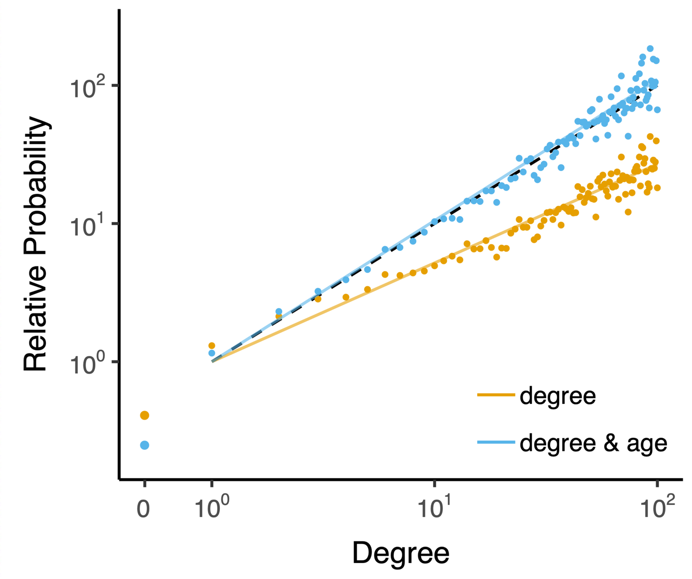
- One can test and compare the likelihood of different formation processes for a specific data set.
- For example, preferential attachment can be hard to distinguish from other processes just from outcome data.
- To illustrate, we generate synthetic data with a process that varies the relative role of degree (p) and triadic closure (r).
- We then estimate the power-law exponent γ to test for the presence of PA in the outcome graphs.



Application – Citation Network

- We apply the logit framework to fit a series of models to a large citation network.
- Here are the resulting regression coefficients (left) and non-parametric estimates for the role of degree in the form of prior citations (right) for two of these models.

log Citations	0.717*	1.052*
	(0.008)	(0.012)
Has degree	1.684*	1.862*
	(0.053)	(0.063)
Has same author		5.928*
		(0.114)
log Age		-1.096*
		(0.018)
Observations	10,000	10,000
Log Likelihood	-20,799	-14,384
Test accuracy	0.358	0.533
Note:		* p<0.01



- Just accounting for degree results in sub-linear preferential attachment, while accounting for age results in linear preferential attachment ($\alpha \approx 1$). The non-parametric estimates are remarkably linear.
- In the paper we also do an analysis with Flickr data.

Future Work

We are currently exploring a number of extensions to this work:

- Stratified negative sampling to improve efficiency
- Node heterogeneity of parameter estimates
- Different processes for choosing i
- Modeling edge deletion
- Other feature parity with SAOM models