Exponential Random Graph Modelling in Learning Analytics

Download Tutorial Materials:
https://github.com/ndowell/LASI17-ERGM-Tutorial

Oleksandra Poquet[1] & Nia Dowell[2]
University of South Australia[1]
University of Michigan[2]

Tweet your love/hate to #lasi2017 or at @choux and @NiaDowell

Quick Agenda

- Introduction: Why ERGMs
- Worked Example
  - Data Exploration
  - Building a Model
  - Evaluating a Model
  - Reporting
- Examples of LA applications

Who we are

Sasha Poquet

Postdoctoral Research Fellow,
School of Education
University of South Australia

Research foci:

- Social aspects of learning (social context indicators)
- Identity and network formation in educational settings
- Relationship between learning design and social context

Nia Dowell

Postdoctoral Research Fellow,
School of Information & DIG
University of Michigan

Research foci:

- Discourse analysis in higher education
- Understanding group interaction dynamics
- Modeling social, cognitive and affective processes

Introductions, why are we here?

Introductions around the room:

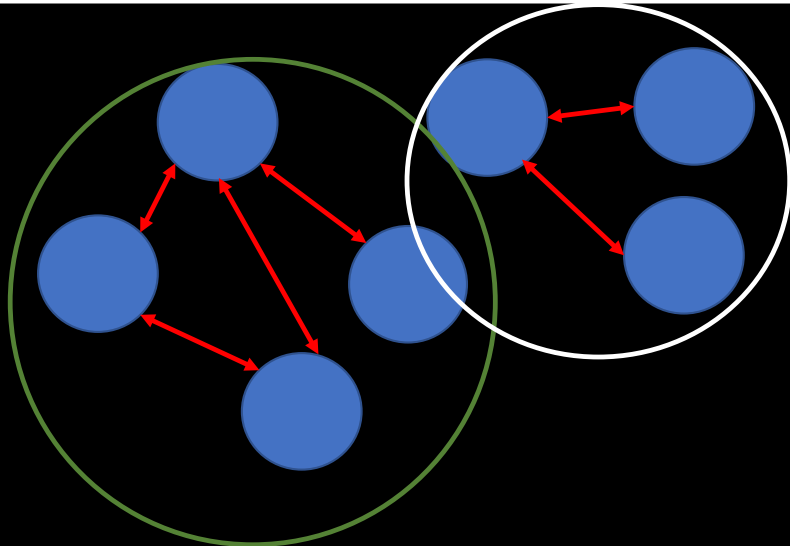
- Name, affiliation, role, a little background
- What drove you to register for this tutorial?
- What is your familiarity with SNA and do you have any particular datasets in mind?

Understand how networks form and evolve

Allowing us model the structural tendencies of network forms on the basis of multiple theories and at multiple levels to understand the driving forces

Briefly about Networks

- Nodes -> people, ideas, artefacts, words
- Links -> type of relation between the entities, ‘likes’, ‘is neighbors with’, ‘co-occurs with’, ‘clicked on’, etc.
- Has boundaries
- Can be represented in matrices



Asks advice from

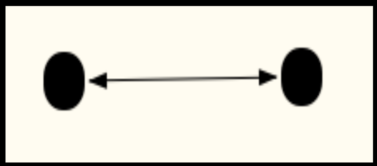
	John	Lisa	Sasha	Nia
John	0	0	1	0
Lisa	0	0	1	1
Sasha	1	1	0	1
Nia	0	1	1	0

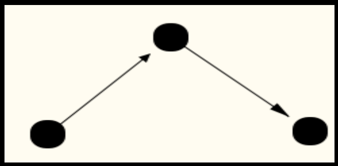
Statistical Analysis of Networks is Fundamentally different

- RELATION is a unit of analysis
  - How many relations in a dataset, not how many nodes.
- RELATIONS can be dependent.
  - Can we truly conclude that such relationship is random : "A friend of a friend is my friend"?

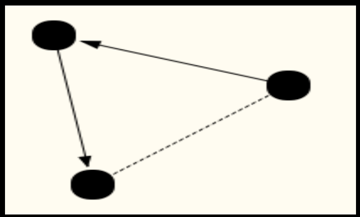
Why do networks form?

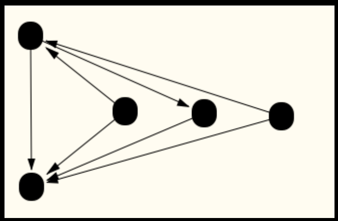
Reciprocity



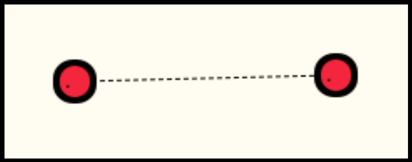


Transitivity





Homophily

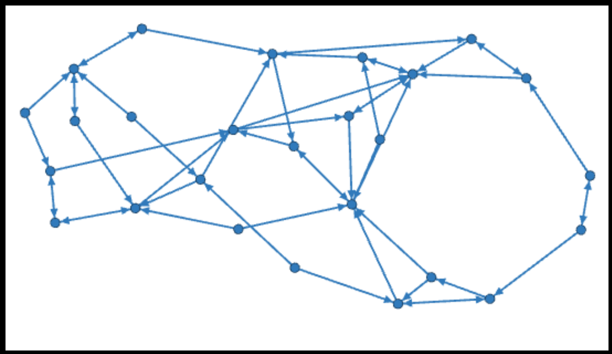


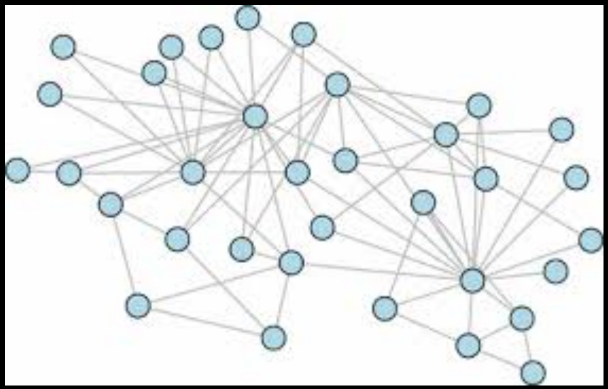
Why do these networks *look* like this?

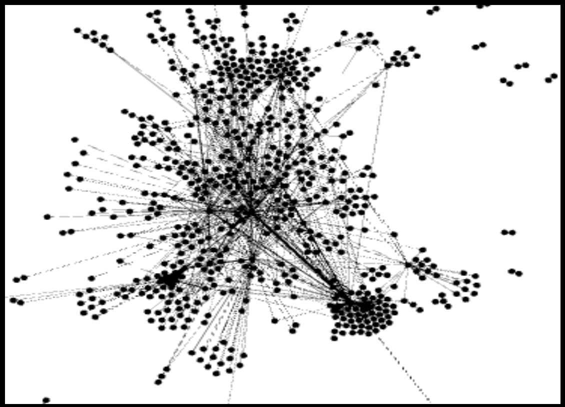
Reciprocity

Transitivity

Homophily





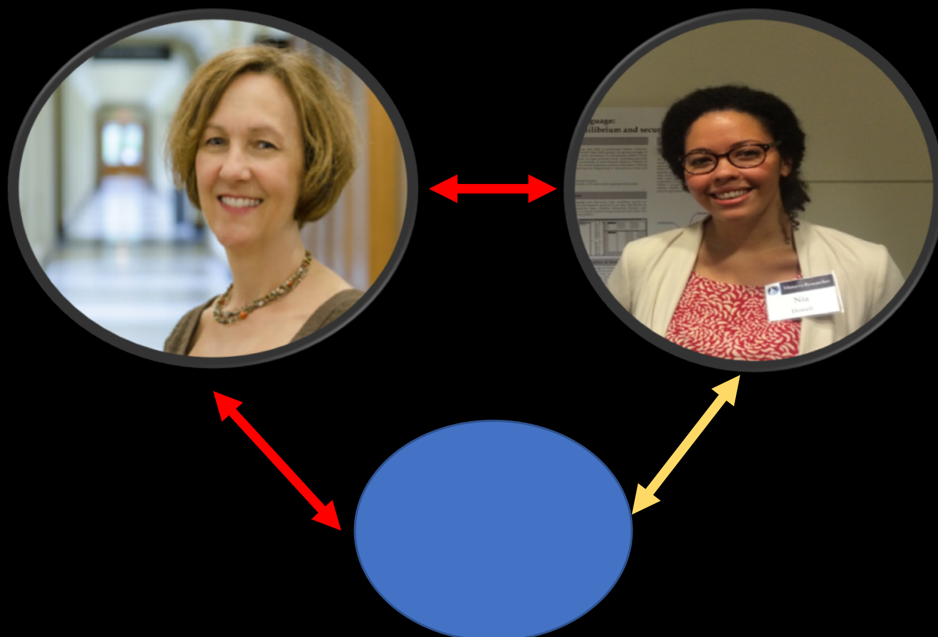


Are some processes more important for network formation than others?

How do node attributes impact tie formation?

Challenge 1

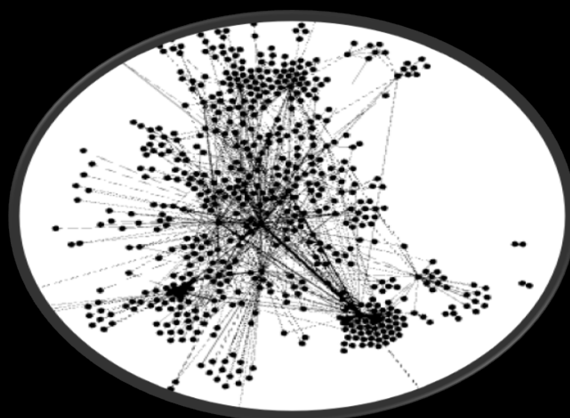
- Modelling Network Ties (with ERGMs) allows
  - To handle non-independence of observations



Challenge 2

Exogenous (external)

- Race
- Gender
- Religious affiliation
- Age



Endogenous (internal)

- Reciprocity
- Transitivity
- Cyclicality

Examples

- Does posting in week N predict posting in week N+1? [Exogenous]
- What is the probability of learners with higher grades to interact with one another? [Exogenous]
- Are learners more likely to answer questions of other learners that previously answered their question? [Endogenous]

Modelling Network Ties (with ERGMs) allows

- To model multiple processes
  - Include multiple parameters in the model like homophily, transactivity, reciprocity. These parameters can be grounded in several theories.

- Theories of self-interest
- Theories of mutual interest and collective action
- Theories of social and resource exchange

- Theories of contagion
- Theories of balance
- Theories of homophily
- Theories of proximity
- Theories of co-evolution

Contractor, Wasserman & Faust, 2006

Social Networks 37 (2014) 1–13

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Relationship patterns in the 19th century: The friendship network in a German boys' school class from 1880 to 1881 revisited

Richard Heidler[a,*], Markus Gamper[b,c], Andreas Herz[d], Florian Eßer[d]

[a] Department for Education and Social Sciences, Bergische University Wuppertal, Gaußstr. 20, 42097 Wuppertal, Germany
[b] Department of Sociology, University of Trier, Cluster of Excellence "Societal Dependencies and Social Networks", Germany
[c] Faculty of Arts and Humanities, University of Cologne, Germany
[d] Institute for Social Pedagogy and Organization Sciences, University of Hildesheim, Lübeckerstraße 3, 31141 Hildesheim, Germany

A R T I C L E I N F O

Keywords:
Friendship network
19th century network data
ERGM
Popularity
School achievement

A B S T R A C T

The article presents a friendship network from 1880 to 1881 in a school class, which goes back to the exceptional mixed-methods study of the German primary school teacher Johannes Delitsch. The re-analysis of the historic network gives insights into what characteristics defined the friendship networks in school classes in Germany at the end of the 19th century. ERGMs of the so far unmarked data show structural patterns of friendship networks similar to today (reciprocity, transitive triadic closure). Moreover we test the influence of the class ranking order (Lokationsprinzip), which allocates the pupils in the class room according to their school performance. This ranking order produces a hierarchy in the popularity of pupils, through hierarchy–congruent friendship ties going upwards in the hierarchy. In this respect, concerning the effect of school achievement on popularity, we find a strong stratification, which is not always prevalent today.

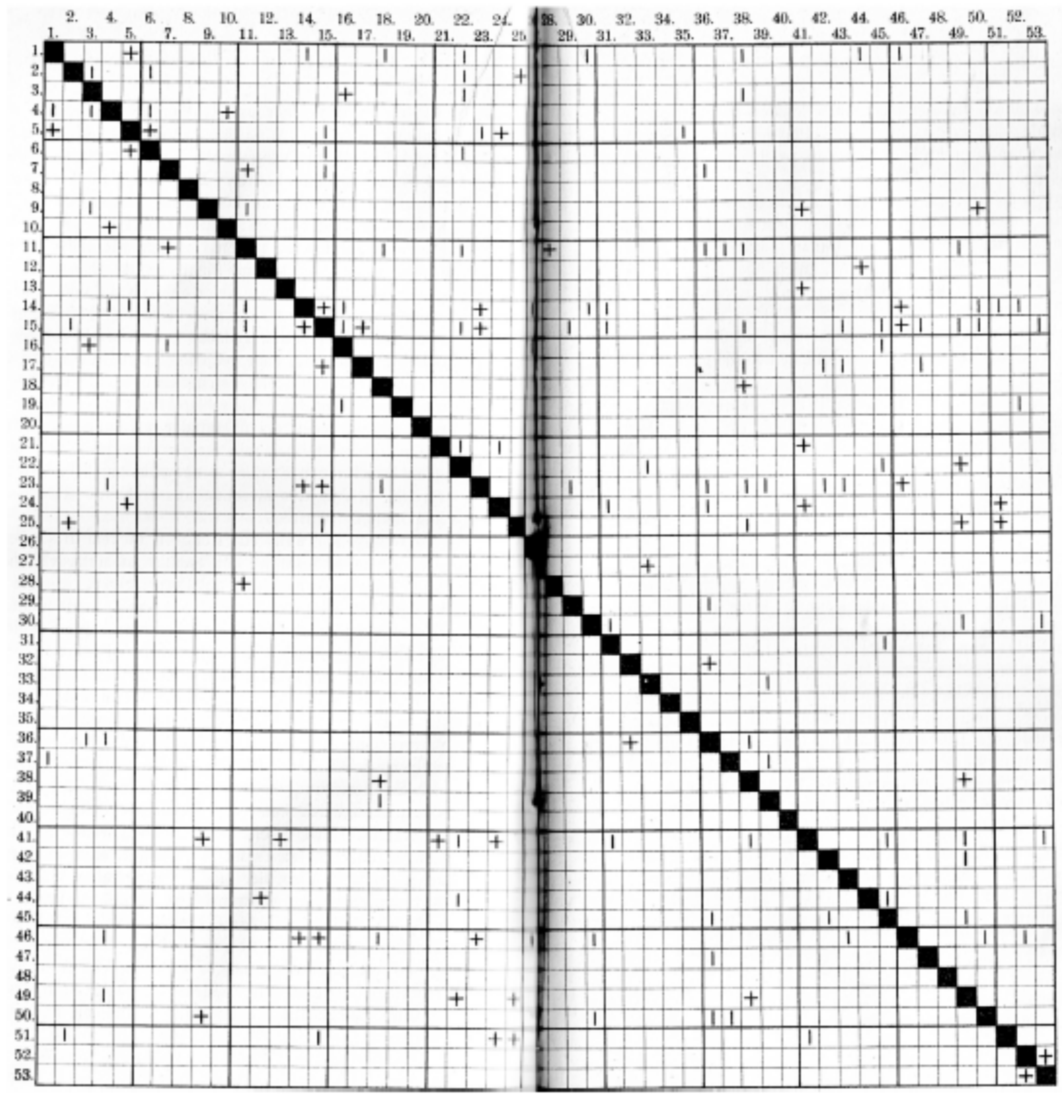


Fig. 2. Original matrix of friendship as it was delineated by Delitsch (1900: 160f).

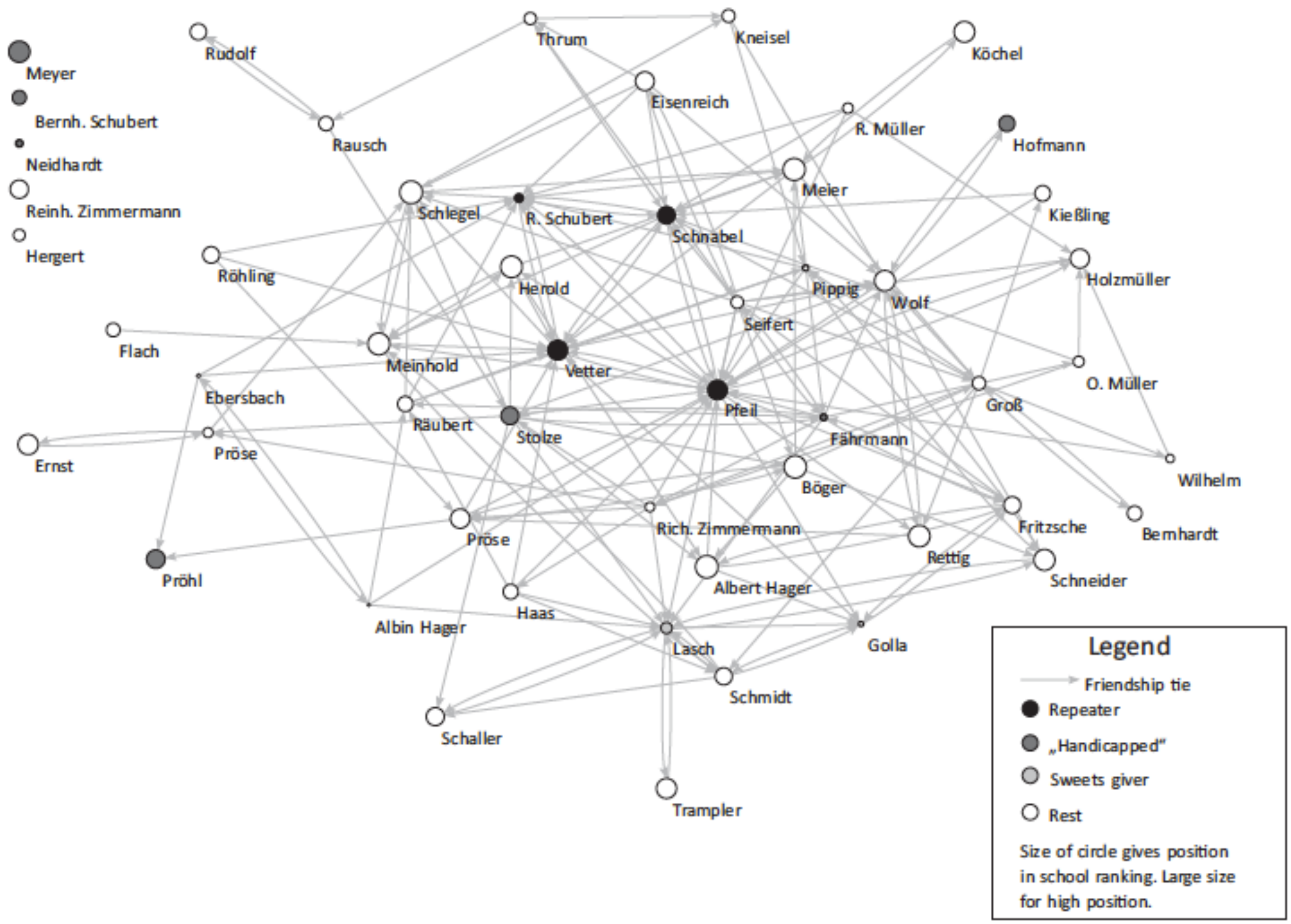


Fig. 3. Visualization of friendship network (Spring-embedder Kamada–Kawai).

Table 3
Results for multivariate ERGM.

	Model 1 (rank-popularity)			Model 2 (up-rank)			Model 3 (rank-popularity + up-rank)		
	Coefficient	Std. error	*p*-Value	Coefficient	Std. error	*p*-Value	Coefficient	Std. error	*p*-Value
Intercept									
Edges	−3.7590	0.3077	0.0000*	−4.1749	0.3983	0.0000*	−4.1283	0.4506	0.0000*
Up-rank									
Up-rank	–	–	–	0.7436	0.2324	0.0014*	0.6280	0.4052	0.1213
Main effects on indegree									
Class ranking (rank-popularity)	0.0157	0.0053	0.0029*	–	–	–	0.0035	0.0086	0.6845
Repeaters and sweets giver	1.1325	0.1963	0.0000*	1.1302	0.1913	0.0000*	1.1405	0.1971	0.0000*
'Handicapped'	−0.4177	0.2996	0.1634	−0.4231	0.3005	0.1593	−0.4192	0.3000	0.1625
Main effects on outdegree									
Class ranking	−0.0092	0.0051	0.0705	0.0060	0.0058	0.3041	0.0033	0.0099	0.7401
Homophily effects									
Class ranking (absolute difference)	−0.0060	0.0062	0.3371	−0.0031	0.0062	0.6137	−0.0037	0.0067	0.5792
Structural effects									
Gwesp (transitivity)	0.7339	0.1413	0.0000*	0.7388	0.1321	0.0000*	0.7326	0.1418	0.0000*
Gwesp (alpha)	0.8818	0.1470	0.0000*	0.8866	0.1467	0.0000*	0.8846	0.1465	0.0000*
Reciprocity	2.5103	0.3618	0.0000*	2.5501	0.3611	0.0000*	2.5465	0.3570	0.0000*
Cyclic triple	−0.8000	0.2827	0.0047*	−0.8026	0.2804	0.0042*	−0.7953	0.2833	0.0050*
Twopath	−0.0149	0.0318	0.6387	−0.0164	0.0298	0.5819	−0.0161	0.0313	0.6064
Null deviance	3821	2756 df		3821	2756 df		3821	2756 df	
Residual deviance	1119	2745 df		1116	2745 df		1116	2744 df	
Deviance	2702	11 df		2705	11 df		2705	12 df	
AIC	1141			1138			1140		
BIC	1206			1203			1211		
MCMC sample size	1,000,000			1,000,000			1,000,000		

* Significance codes: p-value < 0.05.

2. Worked Example

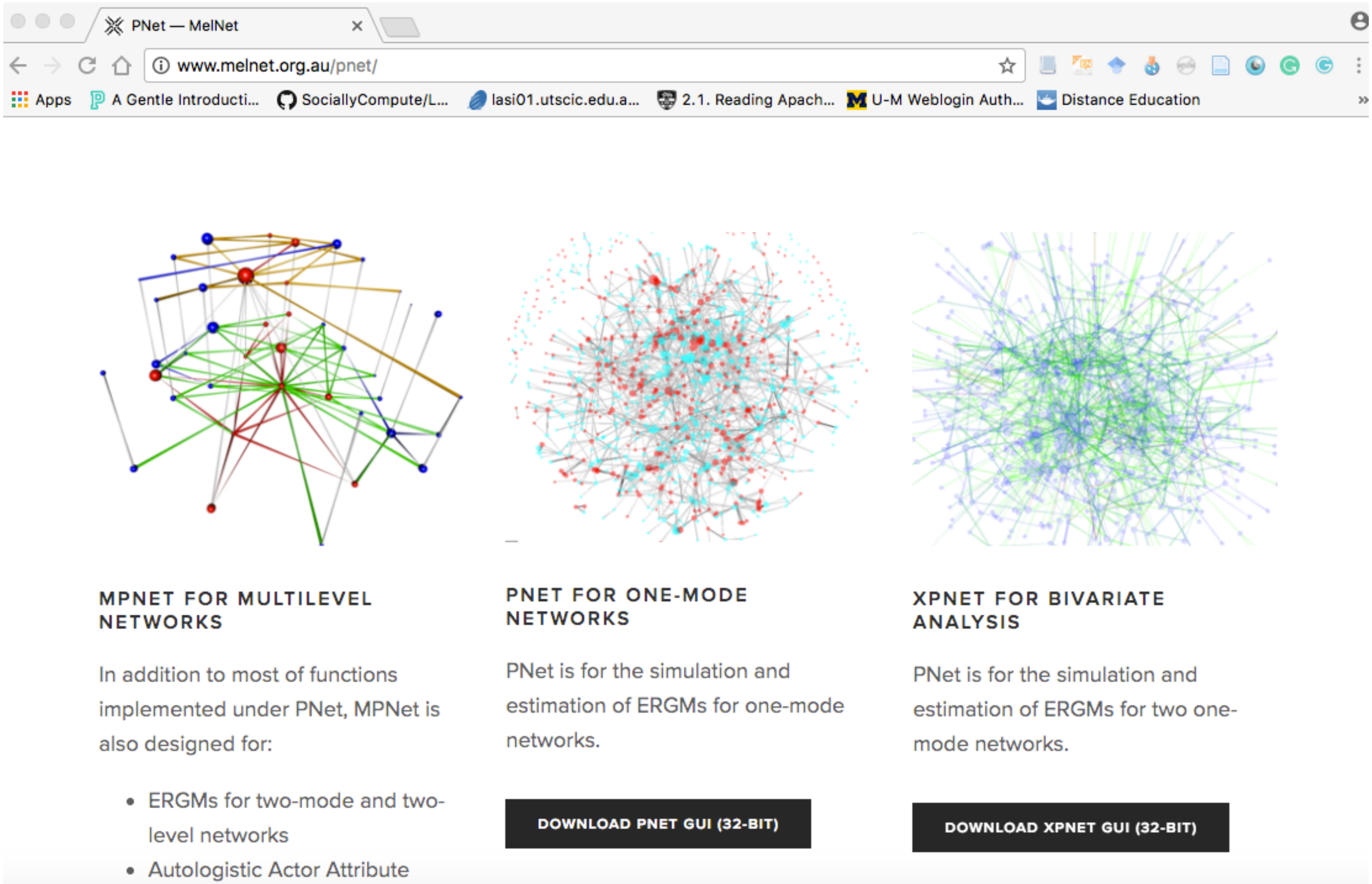
- CCK 11 dataset
- Twitter data
- Included both hashtags and replies
- Current dataset only includes person-to-person interactions
- Demographics was collected manually
- Analysed dataset is an aggregation of 12 week interactions, self-loops removed.

- Skrypnyk, O., Joksimović, S., Kovanović, V., Gašević, D., & Dawson, S. (2014). Roles of course facilitators, learners, and technology in the flow of information of a cMOOC. *The International Review of Research in Open and Distributed Learning*, 16(3).http://dx.doi.org/10.19173/irrodl.v16i3.21702

- Joksimović, S., Kovanović, V., Jovanović, J., Zouaq, A., Gašević, D., & Hatala, M. (2015). What Do cMOOC Participants Talk About in Social Media?: A Topic Analysis of Discourse in a cMOOC. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 156–165). Poughkeepsie, New York: ACM. https://doi.org/10.1145/2723576.2723609

Setting up

- R packages: igraph, ergm, statnet, and intergraph
- Do not load the packages into the workspace yet, i.e. igraph and statnet need to be loaded in different order, and do not work well when loaded at the same time.
- Set the working directory by running 'setwd("your path")'.
- Make sure that the network file g2.gml is in the same folder that you chose as your working directory.
- Start with loading `library(igraph)`
- Remember to `set.seed(234)`



MPNET FOR MULTILEVEL NETWORKS

In addition to most of functions implemented under PNet, MPNet is also designed for:

- ERGMs for two-mode and two-level networks
- Autologistic Actor Attribute

PNET FOR ONE-MODE NETWORKS

PNet is for the simulation and estimation of ERGMs for one-mode networks.

DOWNLOAD PNET GUI (32-BIT)

XPNET FOR BIVARIATE ANALYSIS

PNet is for the simulation and estimation of ERGMs for two one-mode networks.

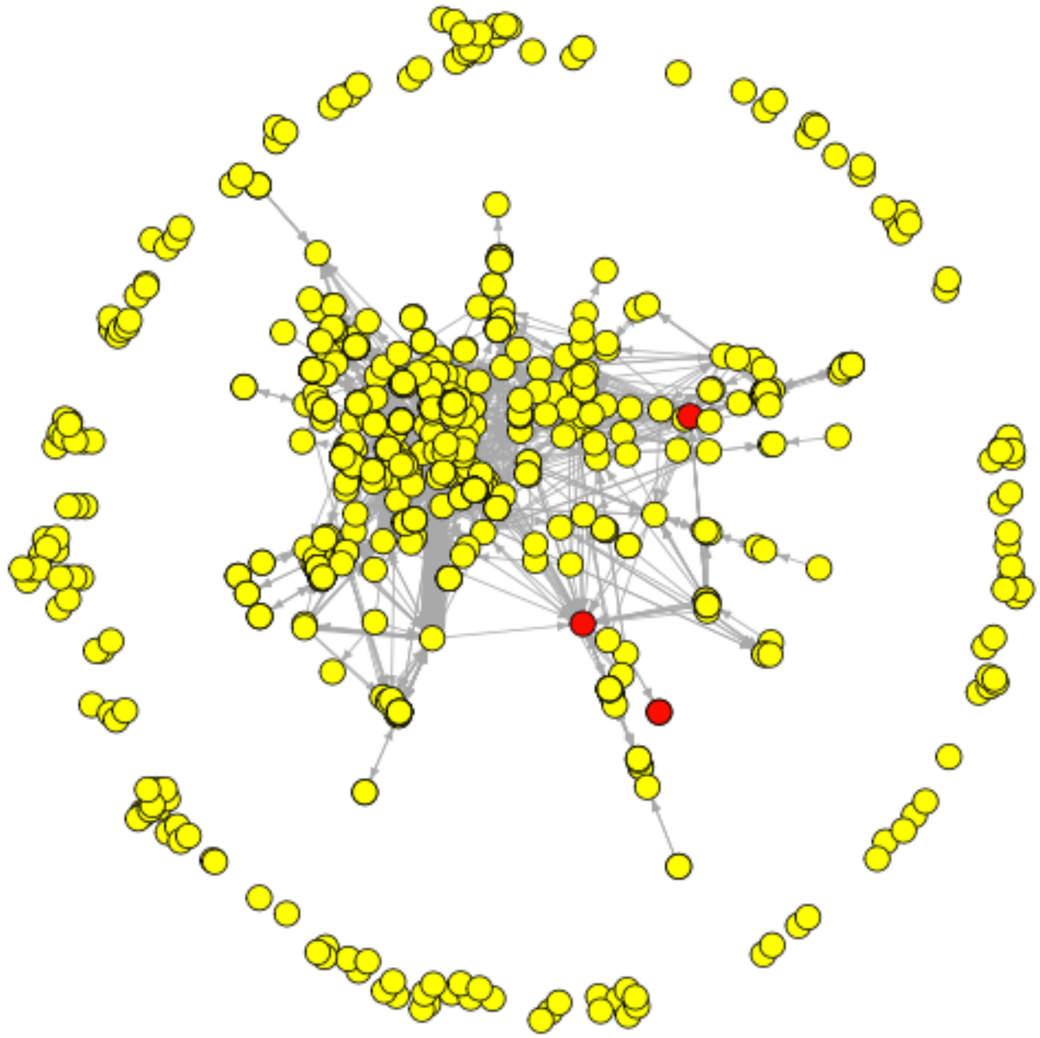
DOWNLOAD XPNET GUI (32-BIT)

Exploring the Data

- Look at the attribute summaries (main effects)
  - Mixing matrices

- Look at the edge weight summaries (for structural properties)
- Look at the degree distribution (for structural properties)
  - Use igraph/statnet/any other

- Look at the counts of features you think you should find within the network
  - Summary (net ~ ergm_term1 + ergm_term2 …)



Attribute Summaries

```
table(V(g)$Continent)
```

```
##
##        Africa              Asia Australia and NZ           Europe
##             2                19               48              191
##  International     North America    South America          Unknown
##             2               232               56              217
```

```
table(V(g)$Gender)
```

```
##
##       F       M     Org Unknown
##     237     285     100     145
```

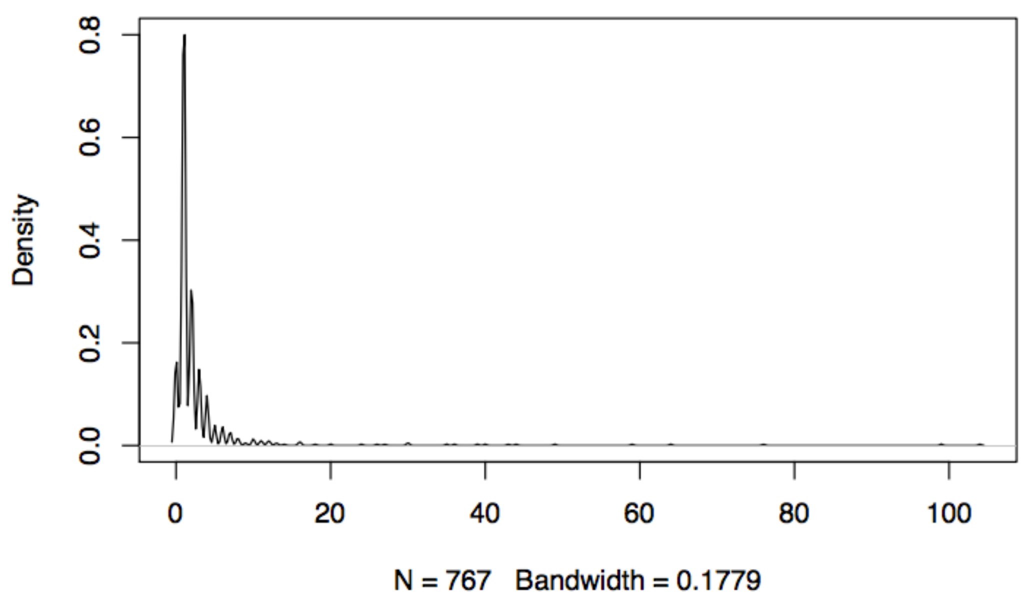
```
table(V(g)$Domain)
```

```
##
##                          Business                        Community
##                                50                                3
## Elementary and primary education                 Entrepreneurship
##                                10                               62
##                        Government                           Health
##                                 4                                9
Higher Education          Languages
             190                 37
         Library       Organization
               6                 68
           Other Secondary education
               9                 78
   Undergraduate            Unknown
               4                216
         Various
              21
```

Network Descriptives

```
summary(E(g)$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.548   1.000  47.000
```



Mixing Matrices

```r
mixingmatrix(net, "Gender")
```

```
##          To
## From        F   M Org Unknown Total
##   F       143 244  67      30   484
##   M       149 240  59      31   479
##   Org      47  69  20      20   156
##   Unknown   9  17   3      45    74
##   Total   348 570 149     126  1193
```

```r
mixingmatrix(net, "Role") #look at the mixing matrices of the network
```

```
##                      To
## From                   Course Instructor Student Total
##   Course Instructor                    1       9    10
##   Student                             98    1085  1183
##   Total                               99    1094  1193
```

```
## From                                 Higher Education Languages Library
##    Business                                        18         8       2
##    Community                                       19         3       0
##    Elementary and primary education                 2         0       0
##    Entrepreneurship                                55         8       0
##    Government                                       0         0       0
##    Health                                           8         1       0
##    Higher Education                               164        20       1
##    Languages                                       28        26       1
##    Library                                          5         0       0
##    Organization                                    39         4       1
##    Other                                            1         0       0
##    Secondary education                             19        16       0
##    Undergraduate                                    0         1       0
##    Unknown                                         30        11       1
##    Various                                         17         2       0
##    Total                                          405       100       6
##                                     To
```

Summary of Structural Features - 1

```
summary(net ~ idegree(0:10)) # what does the in-degree summary for the range of 1 to ten l
```

```
##  idegree0  idegree1  idegree2  idegree3  idegree4  idegree5  idegree6
##       348       259        64        27        18         8         7
##  idegree7  idegree8  idegree9 idegree10
##         8         3         2         6
```

```
summary(net ~ odegree(0:10)) # what does the out-degree summary for the range of 1 to ten
```

```
##  odegree0  odegree1  odegree2  odegree3  odegree4  odegree5  odegree6
##       349       218        99        45        19        12         4
##  odegree7  odegree8  odegree9 odegree10
##         3         2         1         0
```

Summary of Structural Features - 2

```
summary(net ~ edges + mutual
        + triangles + simmelianties
        + intransitive + transitive
        + cyclicalties + twopath)
```

```
##        edges        mutual      triangle simmelianties  intransitive
##         1193            55          1020            18          7415
##   transitive  cyclicalties       twopath
##          964           108          8379
```

Building a model

- Start with Edges = NULL Model
- Structural Features (degree, closure, reciprocity)
- Add main effects
- Add homophily effects
- Put it altogether
- Experiment
- Keep checking for AIC and BIC in the output as you progress in adding features
- Watch out for mixing endogenous with exogenous effects

Null Model – Is the network random?

```
null <- ergm(net ~edges)
```

```
## Evaluating log-likelihood at the estimate.
```

```
summary(null)
```

```
##
## ==========================
## Summary of model fit
## ==========================
##
## Formula:   net ~ edges
##
## Iterations:  8 out of 20
##
## Monte Carlo MLE Results:
##       Estimate Std. Error MCMC % p-value
## edges -6.19741    0.02898      0  <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 814478  on 587522  degrees of freedom
##  Residual Deviance:  17175  on 587521  degrees of freedom
##
## AIC: 17177    BIC: 17189    (Smaller is better.)
```

Controlling for structure (degree, dyads, triads) - 1

```
m1.1 <- ergm(net ~edges + mutual + idegree(2) + odegree(2))
```

```
## Starting maximum likelihood estimation via MCMLE:
## Iteration 1 of at most 20:
## The log-likelihood improved by 3.325
## Iteration 2 of at most 20:
## The log-likelihood improved by 3.267
## Iteration 3 of at most 20:
## The log-likelihood improved by 3.214
## Iteration 4 of at most 20:
## The log-likelihood improved by 2.347
## Iteration 5 of at most 20:
## The log-likelihood improved by 0.06576
## Step length converged once. Increasing MCMC sample size.
## Iteration 6 of at most 20:
## The log-likelihood improved by 0.1261
## Step length converged twice. Stopping.
## Evaluating log-likelihood at the estimate. Using 20 bridges: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
##
## This model was fit using MCMC.  To examine model diagnostics and check for degeneracy, use the mcmc.
```

```
## Formula:   net ~ edges + mutual + idegree(2) + odegree(2)
##
## Iterations:  6 out of 20
##
## Monte Carlo MLE Results:
##          Estimate Std. Error MCMC % p-value
## edges    -6.19795    0.02394      0  <1e-04 ***
## mutual    3.99697    0.15735      1  <1e-04 ***
## idegree2 -1.35739    0.12913      0  <1e-04 ***
## odegree2 -0.85685    0.10860      0  <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 814478  on 587522  degrees of freedom
##  Residual Deviance:  16647  on 587518  degrees of freedom
##
## AIC: 16655    BIC: 16700    (Smaller is better.)
```

Controlling for structure (degree, dyads, triads) - 2

```
m3 <- ergm(net ~edges + mutual + idegree(2) + odegree(2) + triangles)
```

Again, modelling closure with the triangles term was not successful, giving an error message. We can control for both degree and closure using gwesp and gwdegree terms, and adjusting the lamda.

```
m4 <- ergm(net ~ edges + mutual + gwidegree(0.6, fixed=T))
## ===========================
## Summary of model fit
## ===========================
##
## Formula:   net ~ edges + mutual + gwidegree(0.6, fixed = T)
##
## Iterations:  8 out of 20
## Monte Carlo MLE Results:
##           Estimate Std. Error MCMC % p-value
## edges     -5.16486    0.03183      0  <1e-04 ***
## mutual     4.03554    0.16229      1  <1e-04 ***
## gwidegree -3.06987    0.09348      0  <1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      Null Deviance: 814478  on 587522  degrees of freedom
##  Residual Deviance:  15948  on 587519  degrees of freedom
##
## AIC: 15954    BIC: 15988    (Smaller is better.)
```

Controlling for External Effects (1) + Homophily (2)

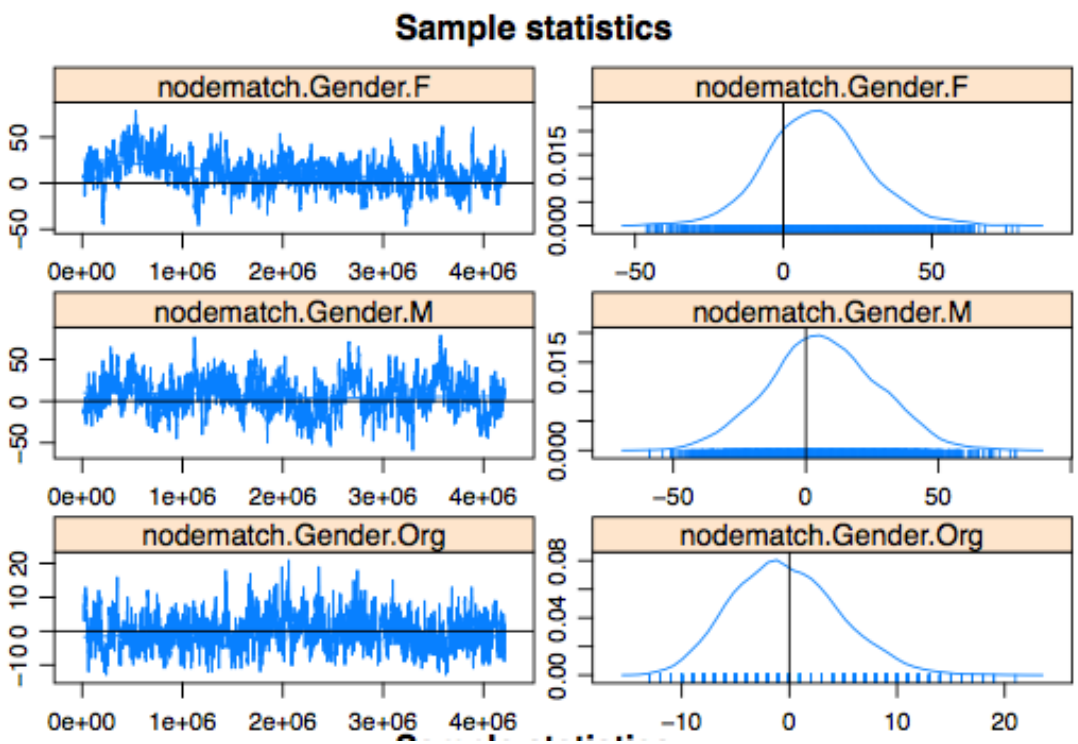
```
## Formula:   net ~ edges + mutual + gwidegree(0.4, fixed = T) + nodematch(
##     diff = T) + nodefactor("Domain")
##
## Iterations:  13 out of 20
##
## Monte Carlo MLE Results:
##                                                Estimate Std. Error  MCMC %  p-value
## edges                                          -5.78473    0.15013       0  < 1e-04 ***
## mutual                                          3.59634    0.16389       1  < 1e-04 ***
## gwidegree                                      -2.53252    0.10763       0  < 1e-04 ***
## nodematch.Gender.F                              0.15082    0.09103       0 0.097541 .
## nodematch.Gender.M                              0.19250    0.07663       0 0.012006 *
## nodematch.Gender.Org                           -0.21463    0.24100       0 0.373155
## nodematch.Gender.Unknown                        1.58599    0.17936       0  < 1e-04 ***
## nodefactor.Domain.Community                     1.35234    0.14653       0  < 1e-04 ***
## nodefactor.Domain.Elementary and primary education -0.12501    0.22240       0 0.574071
## nodefactor.Domain.Entrepreneurship              0.38321    0.08743       0  < 1e-04 ***
## nodefactor.Domain.Government                   -0.33408    0.34008       0 0.325933
## nodefactor.Domain.Health                        0.39091    0.13061       0 0.002763 **
## nodefactor.Domain.Higher Education              0.26964    0.07835       0 0.000579 ***
## nodefactor.Domain.Languages                     0.43595    0.08915       0  < 1e-04 ***
## nodefactor.Domain.Library                      -0.12566    0.29238       0 0.667355
## nodefactor.Domain.Organization                  0.33478    0.09816       0 0.000648 ***
## nodefactor.Domain.Other                        -0.33323    0.25319       0 0.188140
## nodefactor.Domain.Secondary education           0.00479    0.09291       0 0.958885
## nodefactor.Domain.Undergraduate                -1.90561    0.94525       0 0.043802 *
## nodefactor.Domain.Unknown                      -0.41209    0.08974       0  < 1e-04 ***
## nodefactor.Domain.Various                       0.26585    0.10620       0 0.012306 *

## AIC: 15793    BIC: 16030    (Smaller is better.)
```

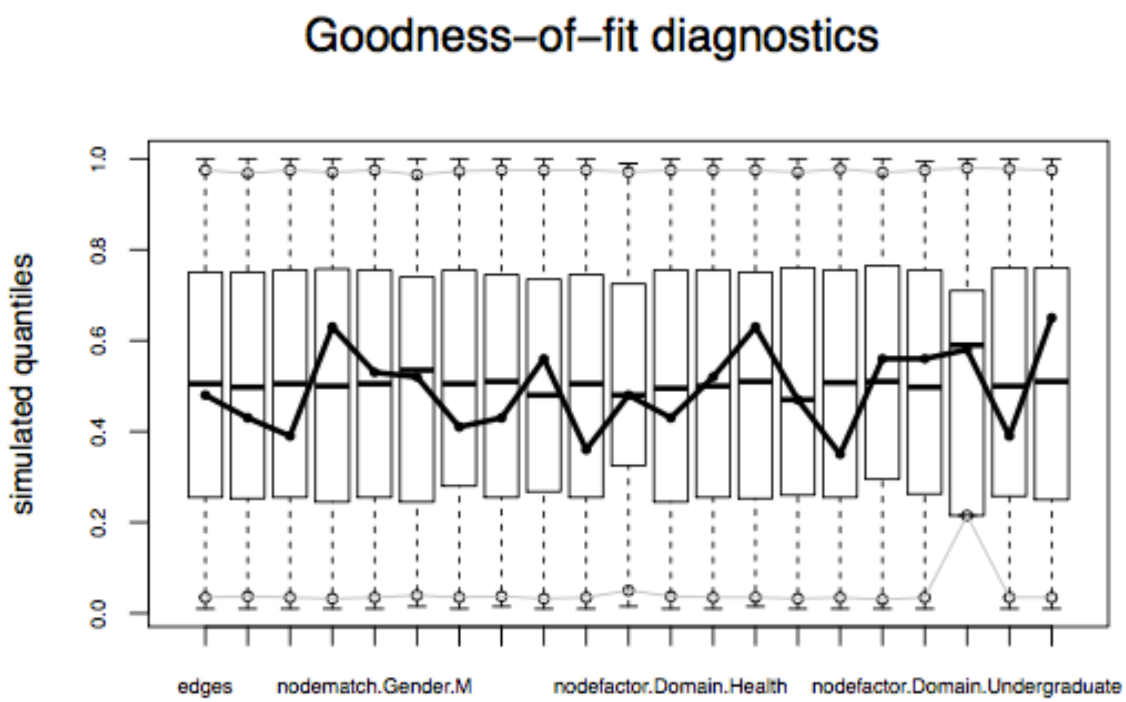
Evaluating the Model

- Has it converged?
- Is AIC/BIC lower than in the null model or previous iterations?
- Does the model resemble the observed network?
- Is there degeneracy in the modelled converged network?

mcmc.diagnostics(m_final)



Observed vs. modelled



What are the odds?

```
lapply(m_final[1],exp) #checking the odds of the data
```

```
## $coef
##                                        edges
##                                   0.00307413
##                                       mutual
##                                  36.46457185
##                                    gwidegree
##                                   0.07945882
##                           nodematch.Gender.F
##                                   1.16278900
##                           nodematch.Gender.M
##                                   1.21227150
##                         nodematch.Gender.Org
##                                   0.80684170
##                     nodematch.Gender.Unknown
##                                   4.88412830
##                  nodefactor.Domain.Community
##                                   3.86645858
## nodefactor.Domain.Elementary and primary education
##                                   0.88249189
##             nodefactor.Domain.Entrepreneurship
```

OK – I did that, and what do I learn?

- Network is described by direct reciprocity, but also some degree of clustering – associated with amplification of information flow and group formation. There is some level of homophily between male participants, community members, entrepreneurs, higher ed, and institutional accounts were instrumental to the communication within the network.

Examples of applications

- SAS data for enrolling students – Who are new students friends with?
- Who speaks to whom – in a variety of settings?
- What context describes the course?
- Etc.

Poquet, O., Dawson, S., & Dowell, N. (2017, March). How effective is your facilitation?: group-level analytics of MOOC forums. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 208-217).

ERGMs outputs for courses with low facilitation

Courses	B	G
Density	**-4.21*** ** (0.03)	**-1.44*** ** (0.3)
Structural Properties		
Reciprocity	**1.29*** ** (0.1)	**1.84*** (0.3)**
Triadic-Level Exchange	**0.56*** ** (0.04)	--
Main Effects		
Moderate Participation	**0.83*** ** (0.02)	--
High Participation	**1.66*** ** (0.05)	--
AIC Null	*38485*	*463*
AIC Final	*32926*	*442*

Interpretation: Network structure similar to informal online communities

Poquet, O., Dawson, S., & Dowell, N. (2017, March). How effective is your facilitation?: group-level analytics of MOOC forums. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 208-217).

ERGMs outputs for courses with high facilitation

Features /Courses	**E**	**I**	**J**
Density	**-3.69***** (0.02)	**-4*****(0.03)	**-3.21*****(0.04)
Structural Properties			
Reciprocity	**0.93***** (0.17)	**1.91*****(0.15)	**0.65***(0.27)
Triadic-level Exchange	**1.13***** (0.08)	**0.31*****(0.06)	**1.05*****(0.13)
Main Effects			
Moderate Participation	**0.53***** (0.023)	**1.33*****(0.04)	**0.73*****(0.03)
High Participation	0.98*** (0.05)	**2.41*****(0.07)	**1.32*****(0.06)
AIC Null	*27098*	*18244*	*18450*
AIC Final	*22005*	*13812*	*14347*

Interpretation: Network structure differs from informal online communities

Where to look for help?

- MelNet Group
- Statnet mailing list
- Ask us:
  - ndowell@umich.edu
  - sasha.poquet@unisa.edu.au