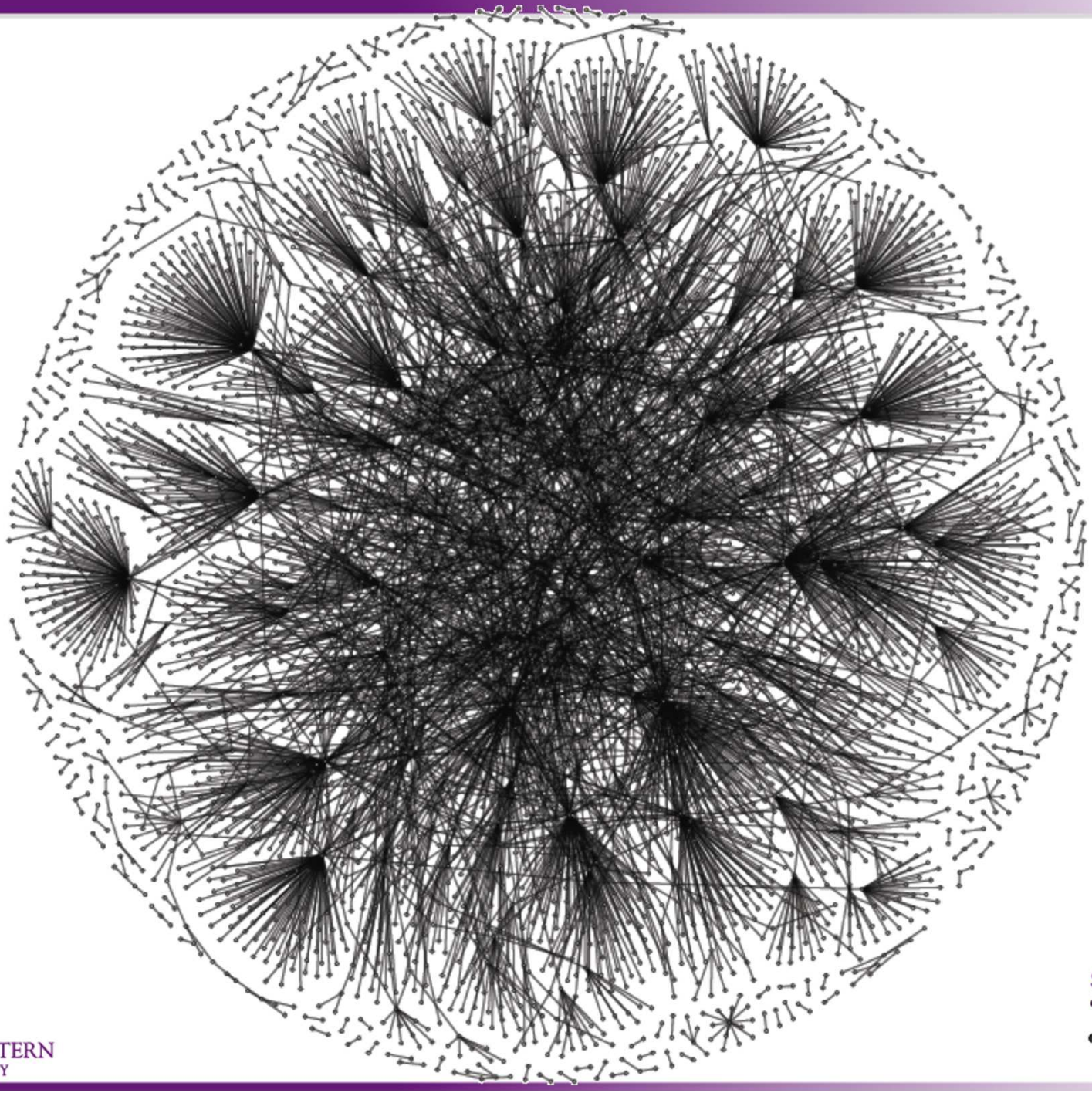
Exploring Twitter Networks in Parallel Computing

Bo Xu[1], Yun Huang[2],and Noshir Contractor[2]

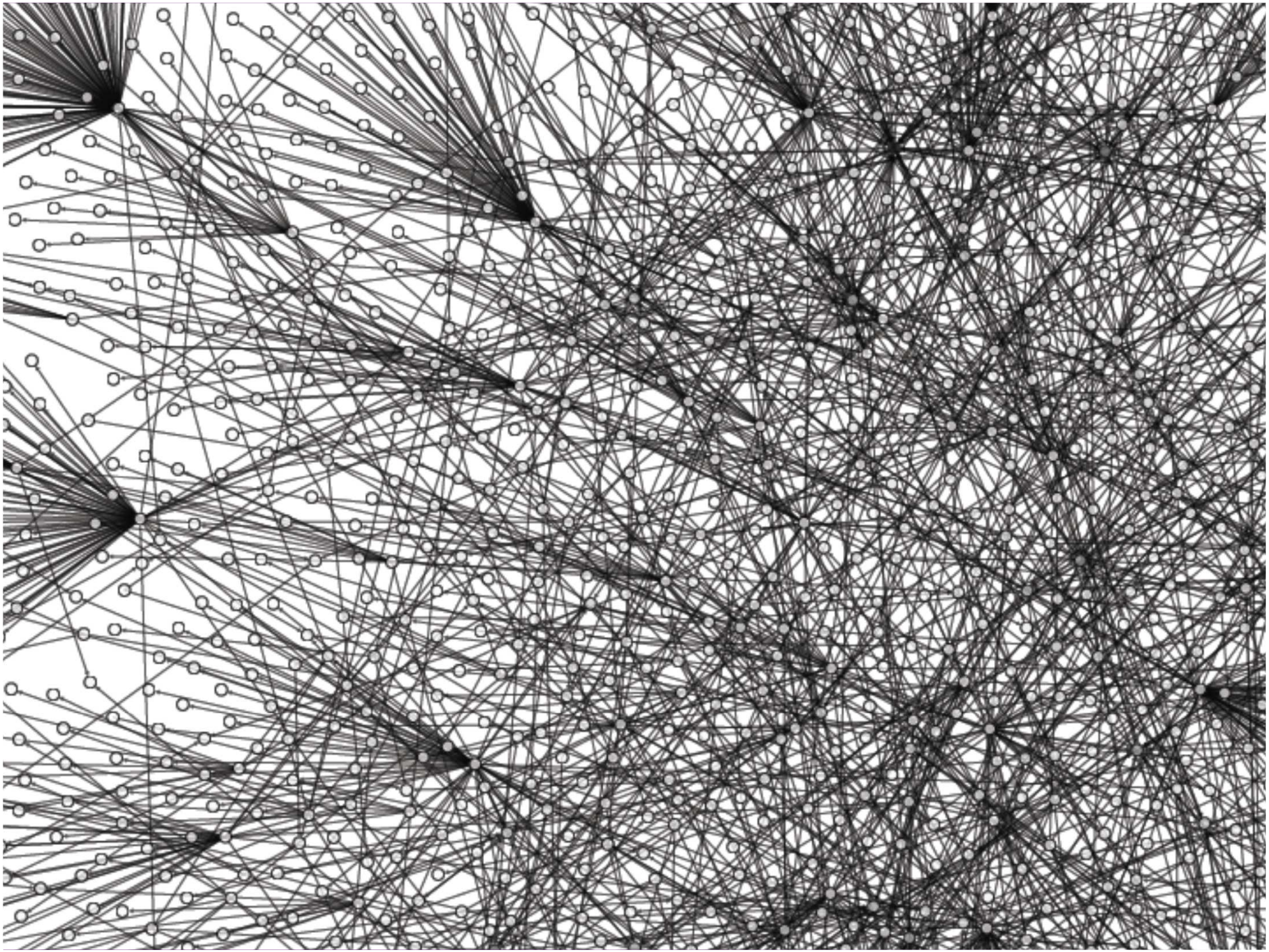
[1] Northeastern University, China
[2] Science of Networks in Communities (SONIC) research group
Northwestern University

This research was supported by grants from the National Science Foundation (Grant No. CNS-1010904, OCI-0904356, OCI-1053575, & IIS-0841583) and Army Research Lab (W911NF-09-02-0053)



NORTHWESTERN
UNIVERSITY

SONIC
advancing the
science of networks in communities



Motivation

- Using parallel computing to explore big networks
  - Visualize network structures of Twitter
  - Quantify the process of link formation
- Media or social networks
  - Heterogeneous nodes and relations of Twitter
- Unfollow behavior
  - Using unfollow to reveal the structure of Twitter networks
  - Identify important factors that influence unfollow behavior using Exponential Random Graph Models

Previous Research on Unfollow

- Autonomous viewpoint
  - Each link in a network is independent of other nodes and links
- Homogenous analysis
  - Performing analysis on entire population
- Findings
  - Both relational and informational factors are important in maintaining relations in Twitter
  - Results do not give much insights on the network

Hypotheses and Network Structures

Relational factors	Informational factors
Reciprocity:	**Informativeness**:
H1. mutual following - → H2. unfollowed by others + →	H6. frequent conversations - →
Social status:	
H3.1. more followers + → / - ← H3.2. more followees - → / + ←	
Social embeddedness:	**Topic homophily**:
H4. share common followees - →	H5. share similar topics - →

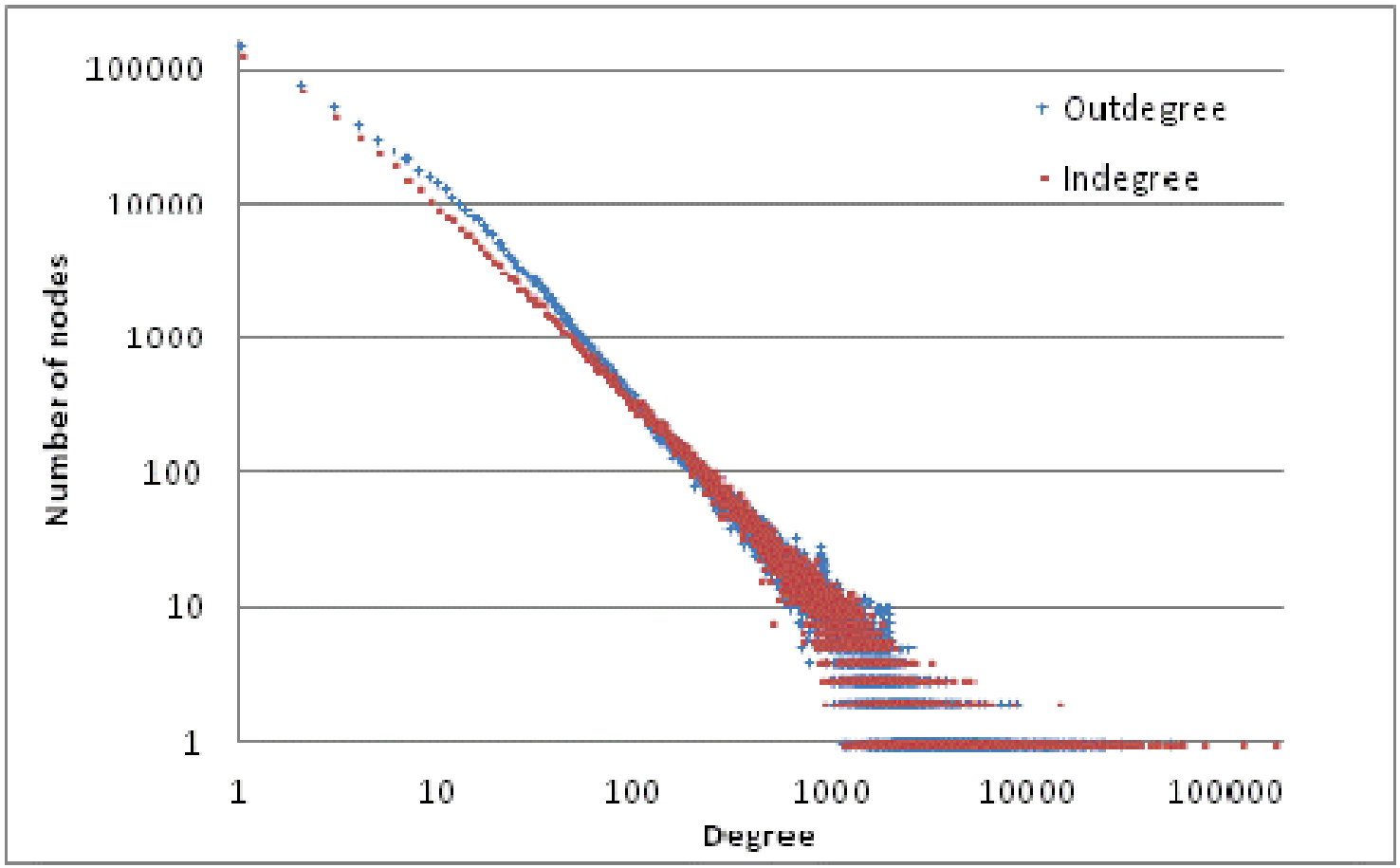
Data Description

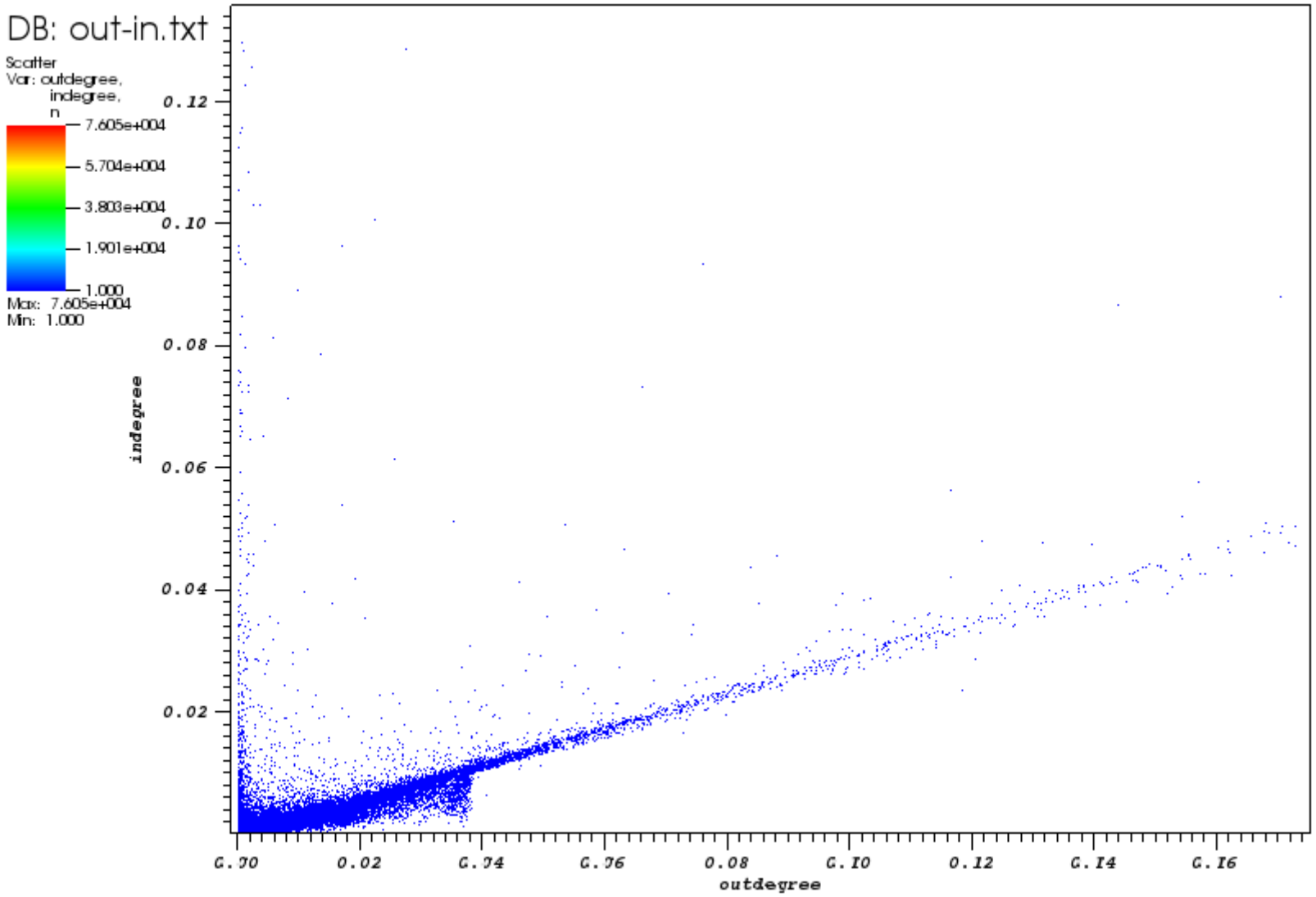
- 697,628 Korean Twitter users w/ two snapshots
  - June 25th and September 3rd 2010
- Construct unfollow networks
  - Dissolution of a tie between snapshots as an unfollow relation: a directed link from user A to user B if A followed B but stopped following at the next time point.

Data Statistics

	Follow network Time 1	Unfollow network Time 2
Vertices	697,628	211,263 involved
Arcs	34,429,170	858,702 unfollow
Mutual pairs	12,676,988 (73.6% of arcs)	33,015 reciprocal unfollow (7.7% of all unfollow ties)

Degree Distributions

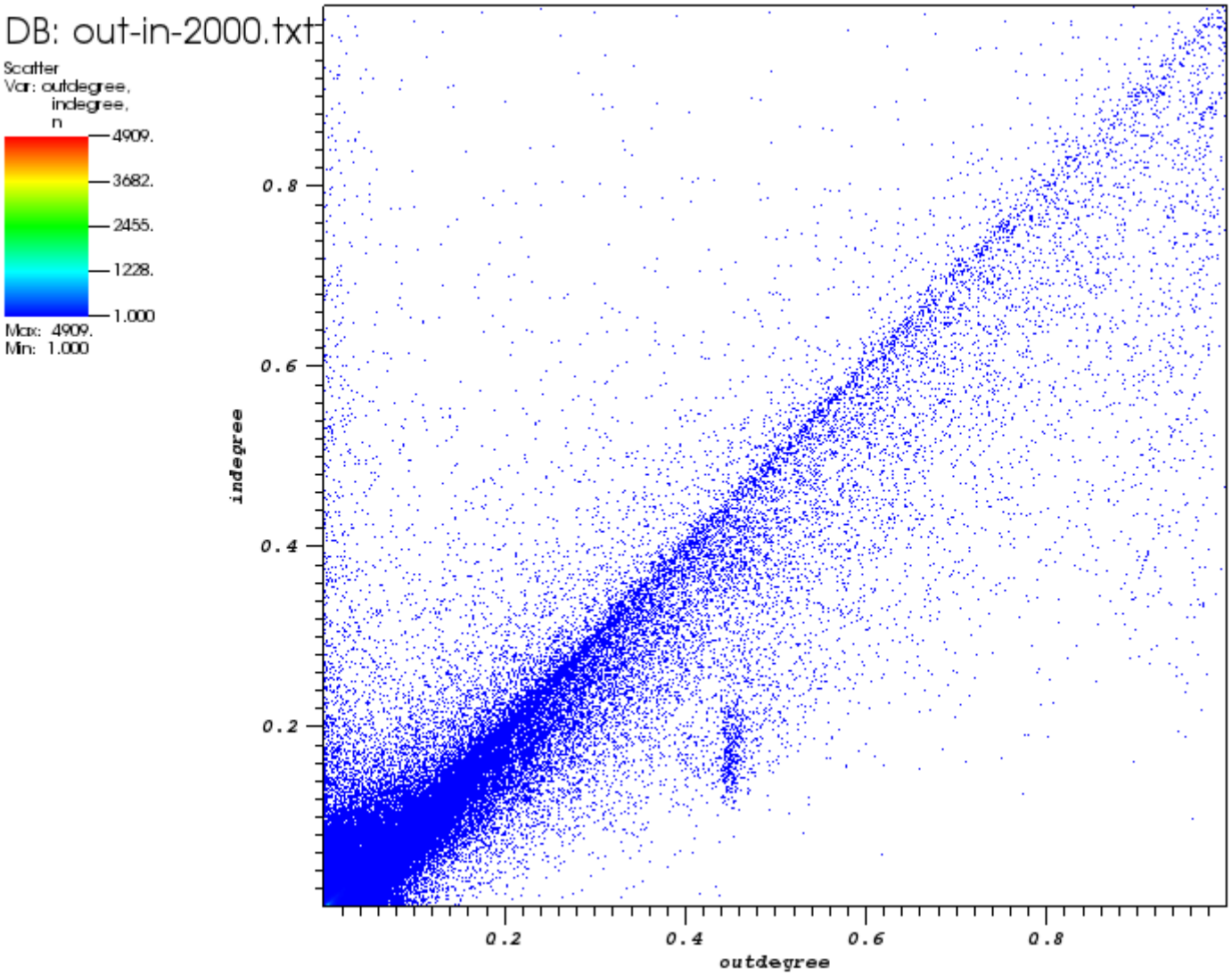




Visualization with VisIt

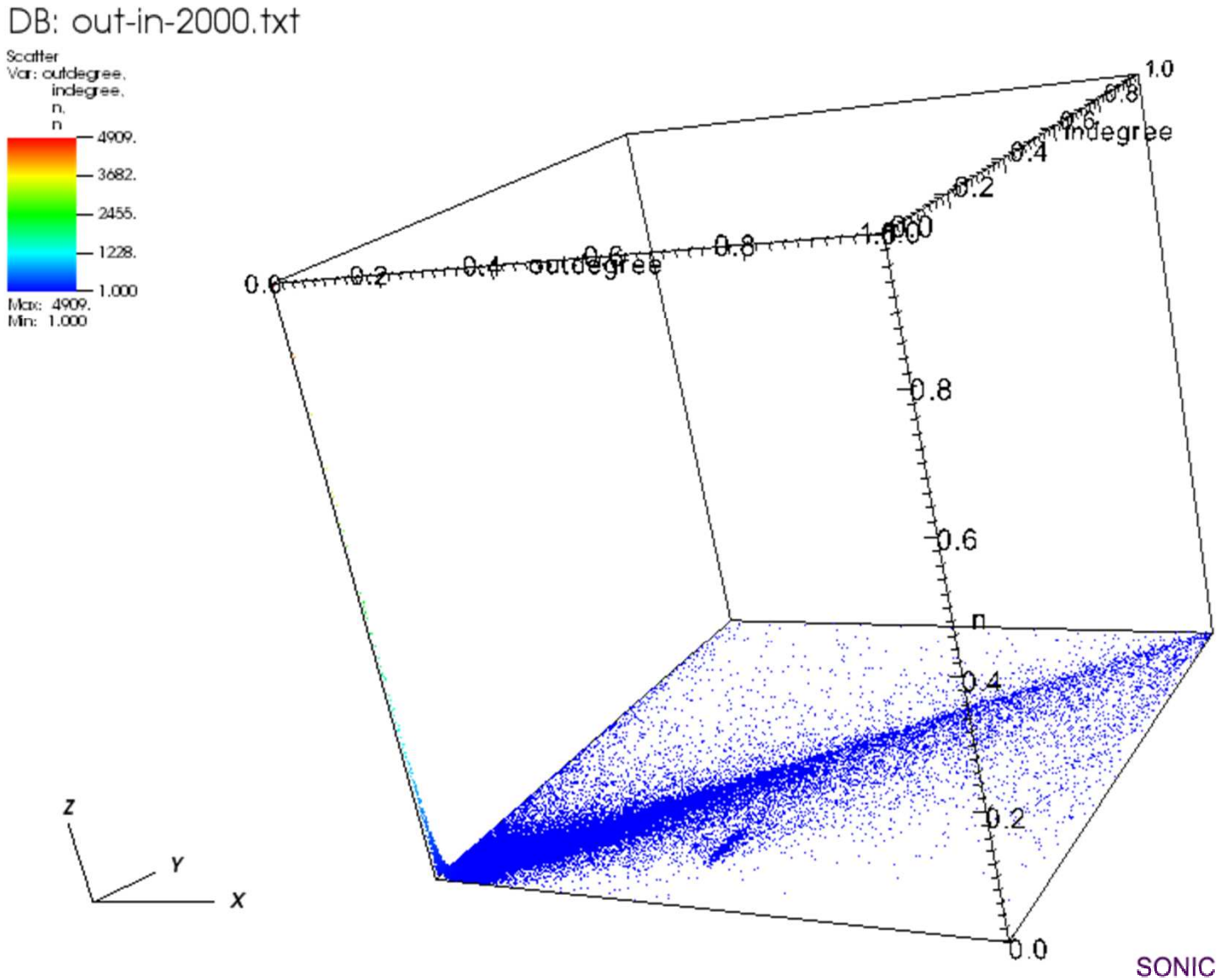
NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities



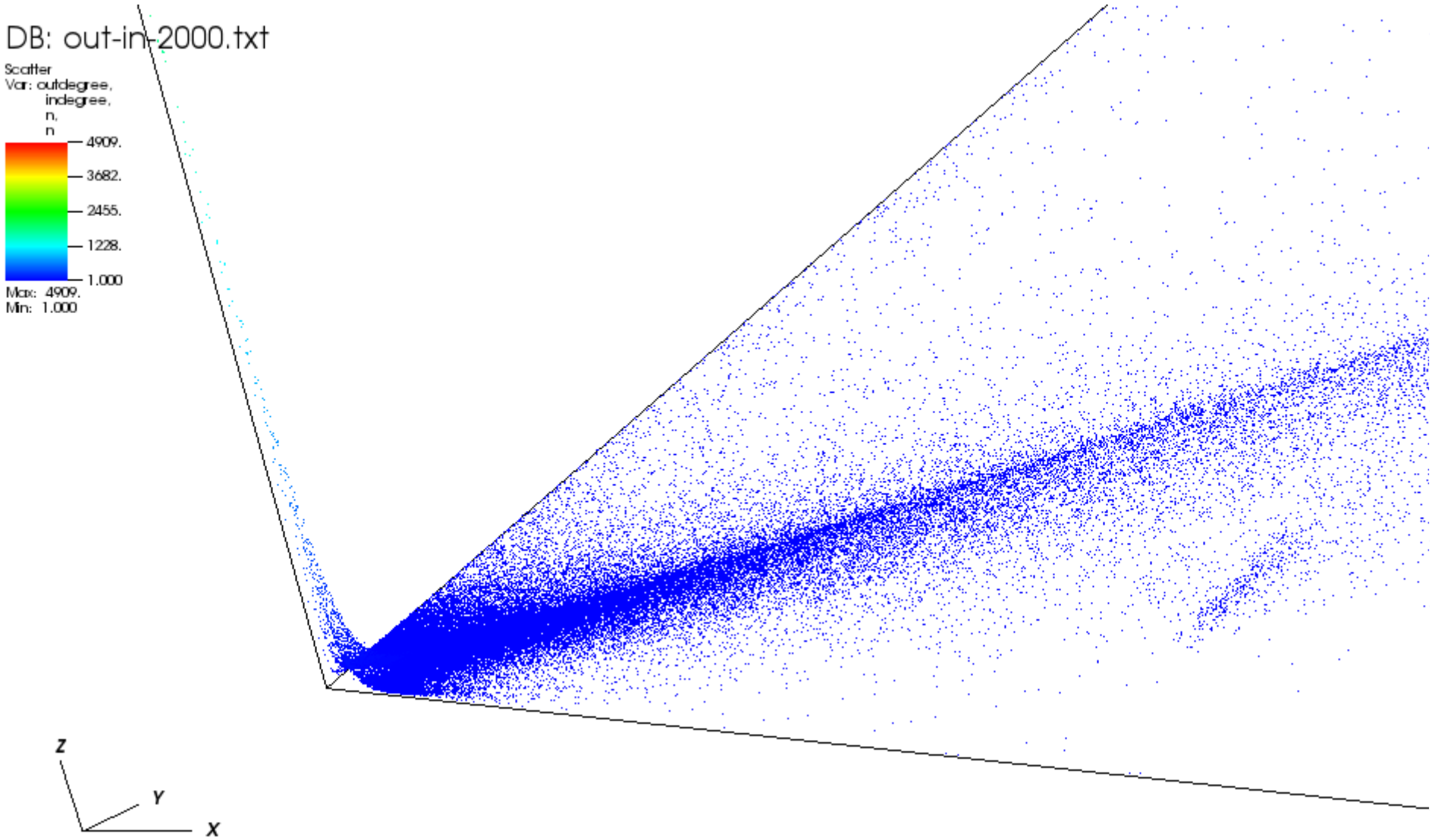
NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities



NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities



NORTHWESTERN UNIVERSITY

SONIC
advancing the
science of networks in communities

ERGM Model

- Interdependent decision to create and dissolve links
- Statistical “MRI” for structural signatures

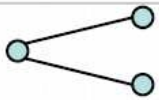
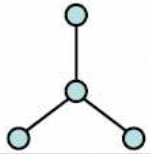
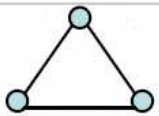
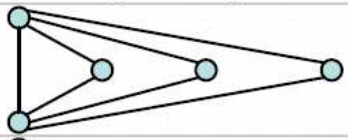
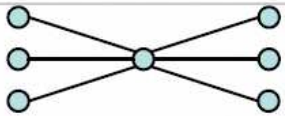
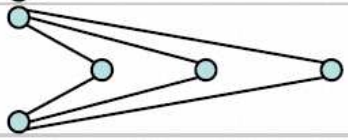
Exponential Random Graph (p*) Model

- ERGMs are a class of stochastic models that share the following general form

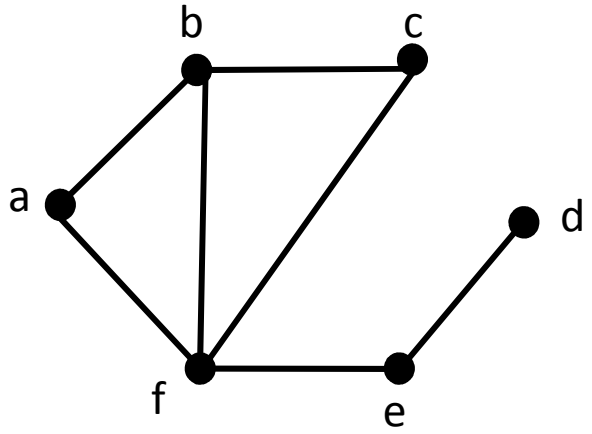
$$P(Y = y) = \frac{1}{k(\theta)} \exp(\theta^T g(y))$$

  - *Y* is the network realization, similar to a random variable
  - *y* is the observed network
  - *g(y)* is a vector of network statistics
  - ϑ^T is a vector of coefficient corresponding to *g(y)*
  - *k(ϑ)* is a normalizing factor calculated by summing up $exp(\vartheta^T g(y))$ over all possible network configurations

Network Statistics (Undirected Networks)

Edge (L)		Isolate	
2-Star (S_2)		3-Star (S_3)	
Triangle (T_1)		Alt-Triangle (AT)	
Alt-Star (AS)		Alt-2-Path (A2P)	

Example:



Edge: 7

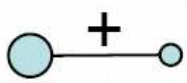
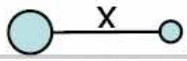
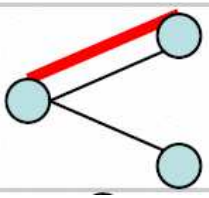
2-Star: 1+3+1+0+1+6=12

3-Star: 0+1+0+0+0+4=5

4-Star: 1

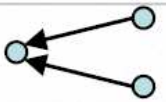
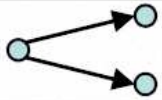
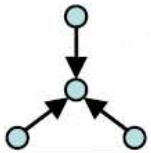
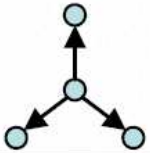
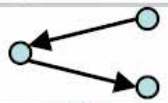
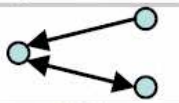
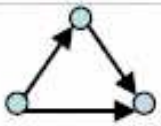
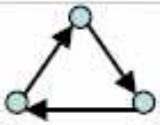
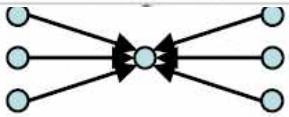
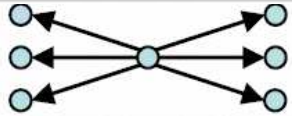
Triangle: 2

Attribute Statistics (Undirected Networks)

● – actors with attribute ○ – actors with or without attribute [Attr] – attribute name			
[Attr]-interaction		[Attr]-activity	
[Attr]-Sum		[Attr]-difference[1]	
[Attr]-interaction			
Dyadic covariate			
[Attr]-Edge		[Attr]-S21	

For more statistics, check Pnet manual:
http://www.sna.unimelb.edu.au/pnet/download/PNet/PNetManual.pdf

Network Statistics (Directed Networks)

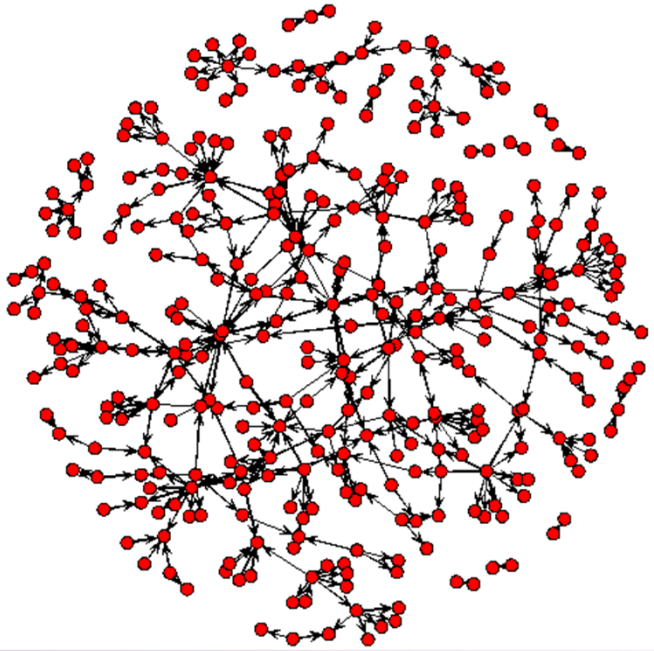
Arc		Reciprocity	
sink		source	
In-2-star		Out-2-star	
In-3-star		Out-3-star	
2-path		T_7	
Transitive Triad (T_9)		Cyclic Triad (T_{10})	
Alt-in-star (AinS)		Alt-out-star (AoutS)	

ERGM Model

- Interdependent decision to create and dissolve links
- Statistical “MRI” for structural signatures
- MCMC simulation
  - Time and space complexity
  - Based on global network statistics (i.e. cannot decompose into local tasks)
  - Inherently sequential

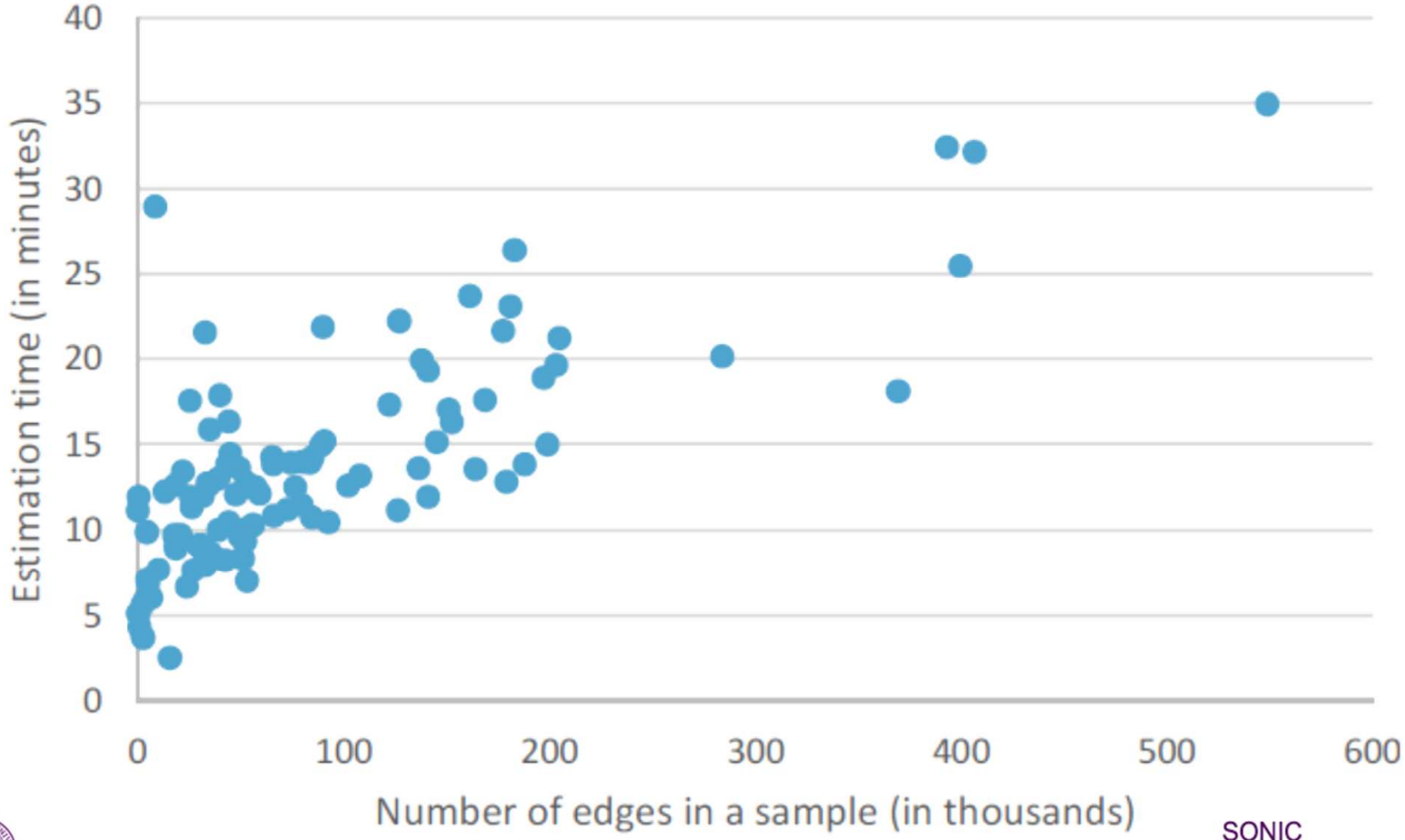
One wave Snowball Sample

1. Randomly select a user with 1000 to 2000 followers as the seed and find all his/her followers in the first snapshot.
2. Extract unfollow relations among these users



Parameters	Estimate (S.E.)	
Unfollow network structures		
Reciprocal unfollow	**3.45***	H2: Supported
Sender attributes of unfollow relations:		
# Followers	**0.21***	H3.1:Supported
# Followees	**-0.47***	H3.2:Supported
Receiver attributes of unfollow relations:		
# Followers	**-0.03***	H3.1:Supported
# Followees	**0.07***	H3.2:Supported
Community in follow networks at time 1:		
Mutual following ties	**-0.46*(.038)**	H1: Supported
# Common followees	**-1.83***	H4:Supported
# Common Hashtag	-0.03	H5:Not supported
Sender's interactions to receiver:		
# Replies	0.001	H6: Not supported
# Retweets	-0.009	H6: Not supported
# Mentions	0.321	H6: Not supported
# Favorite	-0.003	H6: Not supported
Unfollow network structures as control variables:		
Edges / density	**-1.30***	
Weighted out-stars	**-0.57***	
Weighted in-stars	**0.19***	

Performance of Estimation in R/Statnet

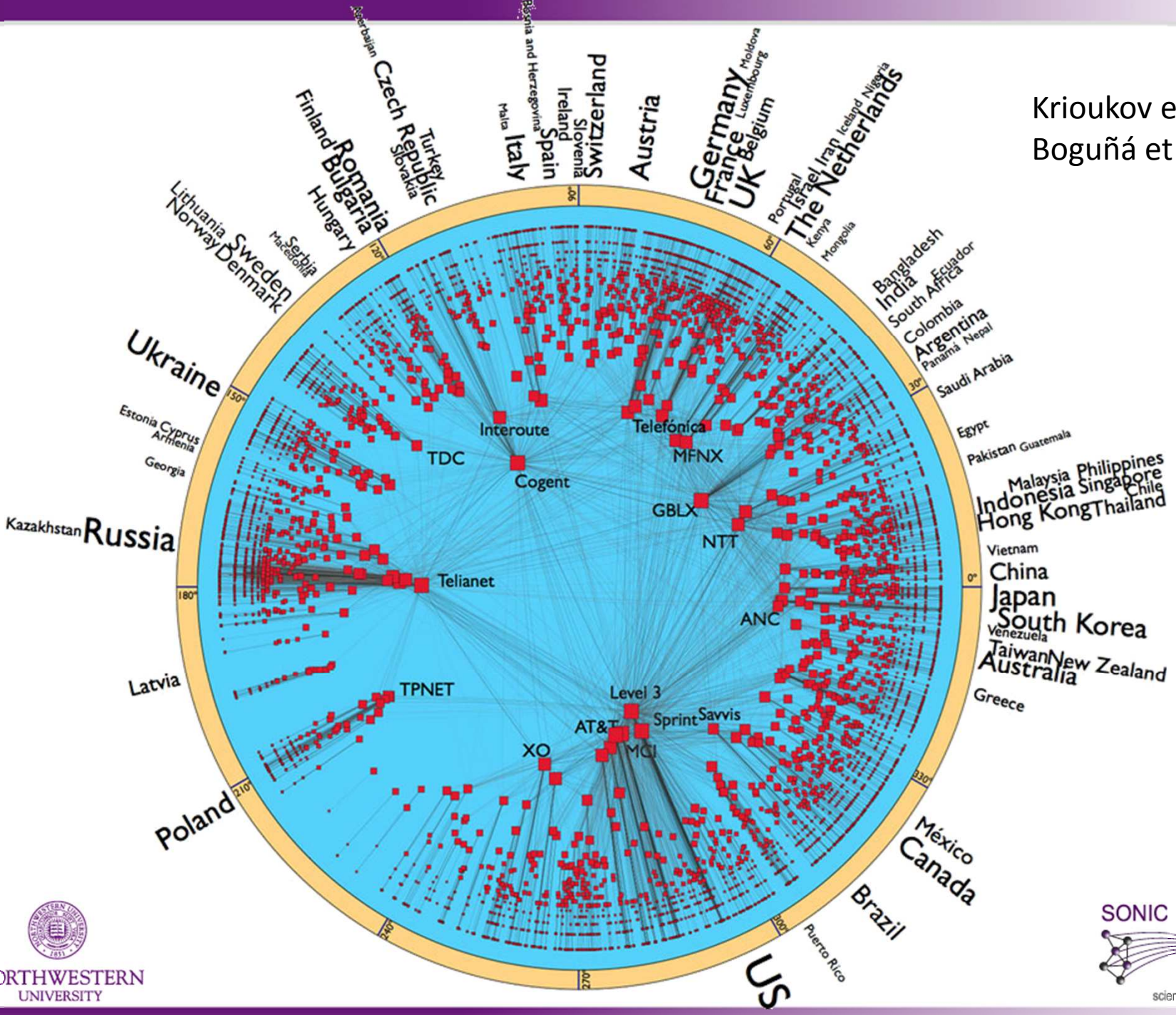


NORTHWESTERN UNIVERSITY

SONIC
advancing the science of networks in communities

Performance of ERGM on Gordon

- Difficulty in parallelization
  - Need of global information in the whole network
  - Sequential update in MCMC simulation
- Snowball sampling and meta-analysis
  - Reveal local structures in large networks
  - Utilize the parallel system as a microscope
  - Limitation of snowball sampling
- Future approach
  - Construct locality in a network
  - Map networks to a hyperbolic space


Krioukov et al. 2010
Boguñá et al. 2010
Azerbaijan
Czech Republic
Turkey
Slovakia
Finland
Romania
Bulgaria
Hungary
Bosnia and Herzegovina
Malta
Italy
Spain
Ireland
Slovenia
Switzerland
Austria
Germany
Moldova
Luxembourg
France
UK
Belgium
Portugal
Israel
Iran
Iceland
Nigeria
The Netherlands
Kenya
Mongolia
Bangladesh
India
Ecuador
South Africa
Colombia
Argentina
Panama
Nepal
Saudi Arabia
Egypt
Pakistan
Guatemala
Malaysia
Philippines
Indonesia
Singapore
Chile
Hong Kong
Thailand
Vietnam
China
Japan
South Korea
Venezuela
Taiwan
New Zealand
Australia
Greece
México
Canada
Brazil
Puerto Rico
US
Lithuania
Norway
Sweden
Denmark
Serbia
Macedonia
Ukraine
Estonia
Cyprus
Armenia
Georgia
Kazakhstan
Russia
Latvia
Poland
Interoute
TDC
Cogent
Telefónica
MFNX
GBLX
NTT
Telianet
ANC
TPNET
Level 3
Sprint
Savvis
AT&T
MCI
XO
0°
30°
60°
90°
120°
150°
180°
210°
240°
270°
300°
330°
NORTHWESTERN UNIVERSITY
SONIC
advancing the science of networks in communities

Hyperbolic Mapping

- Simplify complex networks
  - From graphs to gyrovector space
  - From edges to distances
  - Foundation of task decomposition
- Incorporate heterogeneous structures
  - Detect and visualize communities and structures in complex networks.
- Simplify statistical methods
  - Convert exponential random graphs to auxiliary fields