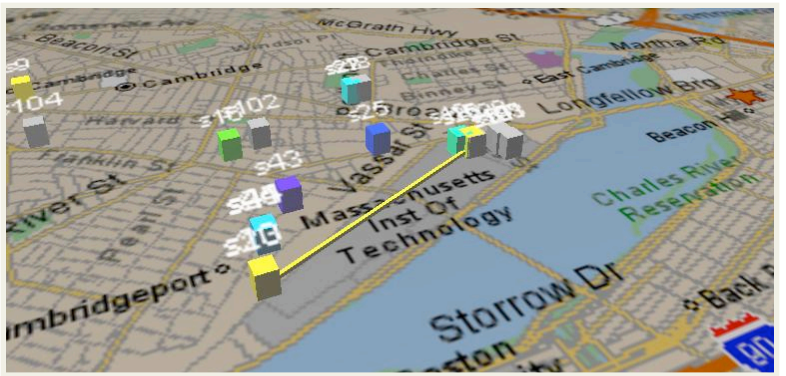
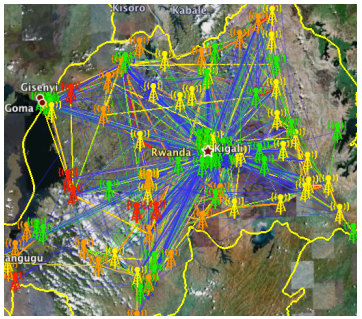


The Scaling of Temporal Human Behavior

… from hundreds, to thousands, to millions, to (Thursday) billions.



Summary of Some Open Questions…

- N = 1 HUNDRED
  - How to infer a relationships from many other temporal behavioral networks?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
  - How to identify the type of edge based on thousands of contextually labeled data points?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
  - How to leverage random sampling to learn about demographic groups?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
  - How to disambiguate spread over a lattice with background prevalence?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
  - How is recent urbanization affecting people's support networks?
  - How can we better understand disease dynamics with actual mobility patterns?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
- N = 10 MILLION
  - How do resources flow through social networks?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
- N = 10 MILLION
- N = 100 MILLION
  - What is driving the behavior of the aggregate?
  - Strength of weak ties?
  - Graph Traversal Using Parallel Binary Search on Sorted Edge Lists?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
- N = 10 MILLION
- N = 100 MILLION
- N = BILLIONS…
  - What does it mean to aggregate this data?
  - How can we do it in a way that improves the quality of life of our species?

Collaborators Welcome

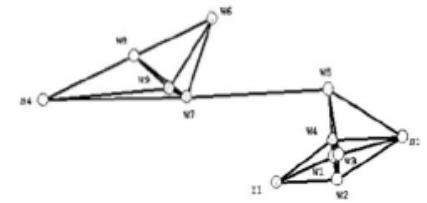
- Aaron Clauset (The Santa Fe Institute)
  Stephen Guerin (RedFish)
  David Lazer (Harvard)
  Mika Raento (Google)
  Hannu Verkasalo (HUT)
  Peter Wagacha (The Univeristy of Nairobi)

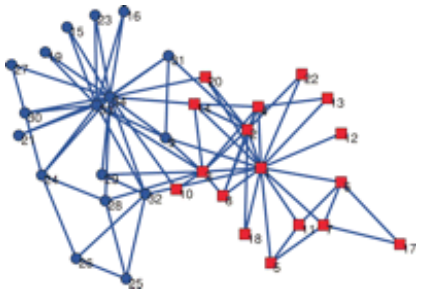
John Quinn (Makerere)
Cosma Shalizi (CMU)
Alessandro Rinaldo (CMU)
Raja Hafiz (CMU)
Caroline Buckee (The Santa Fe Institute / Oxford)
Sune Lehmann (NorthEastern)
Sir Lord Robert May (Oxford)
Yves-Alexandre de Montjoye (Louvain)

Marta Gonzales (NorthEastern)
Dirk Brockmann (NorthWestern)
Marcel Fafchamps (Oxford)
Edo Airoldi (Harvard)
Neil Ferguson (Imperial)
Joshua Blumenstock (Berkeley)
Rob Claxton (British Telecom)
Michael Macy (Cornell)

What We're Comfortable With

Static, Small Graphs



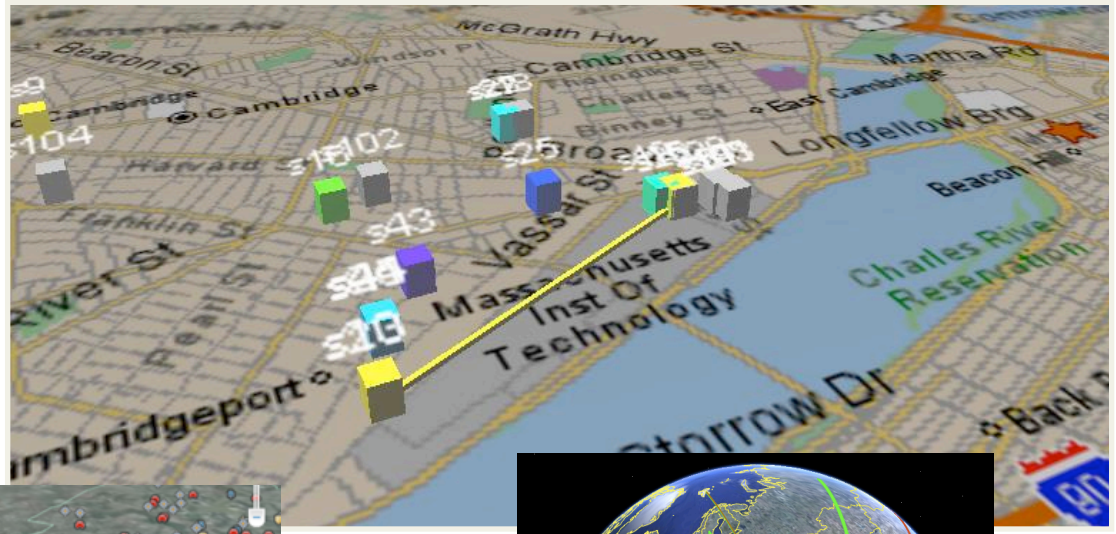


Roethlisberger, F.J., and Dickson, W.J. (1939), *Management and the Worker*, Cambridge, MA: Harvard University Press.

W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* 33, 452-473 (1977).

What We're Not So Good At

Continuous, Weighted, Large Graphs with Dynamic Covariates & **OUTCOMES**.



What We're Not So Good At

Continuous, Weighted, Large Graphs with Dynamic Covariates & **OUTCOMES**.

Talk Take-Away

A Shift in Focus is needed.

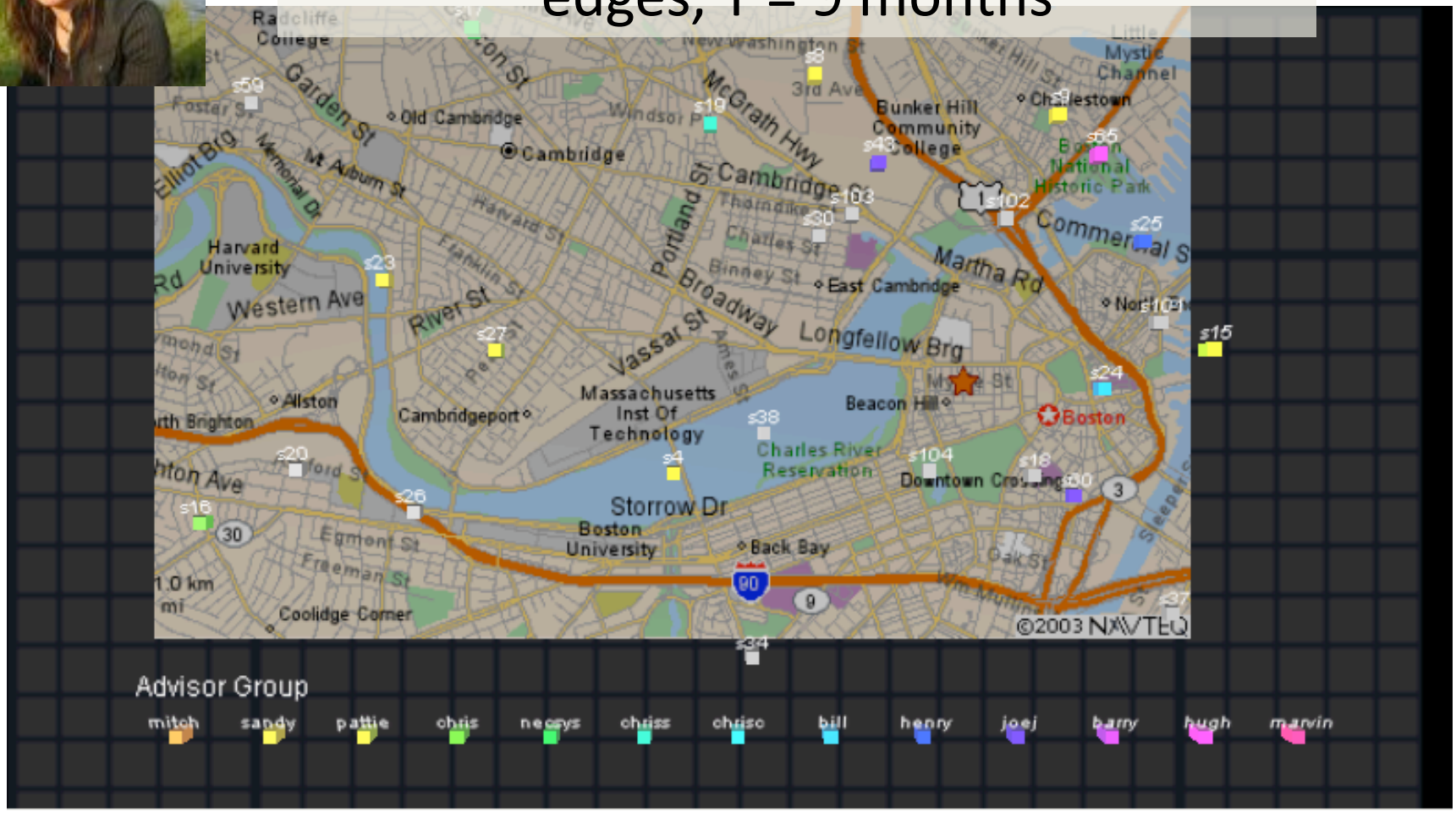
AWAY from distributions and models governing topology

TOWARDS distributions and models for OUTCOMES *informed by* (conditioned on) network topology.

Part 1: 10^2 - 10^5

N = 1 HUNDRED

MIT: 63 relationships, 1.3M proximity edges, T = 9 months



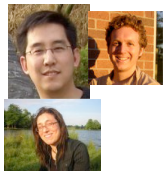
Eagle, N. "Machine Perception and Learning of Complex Social Systems", *PhD Thesis, Massachusetts Institute of Technology*. 2005.

Reality Mining Data

- 100 Nokia 6600s with Context logging software
  - **Location:** Celltower ID / User-Defined Names
    date, area, cell, network, name
  - **Bluetooth:** Proximate Bluetooth Devices every 5 minutes
    date, MAC, device name, device type
  - **Communication:** Phone Call/Text Log
    date, text/call, incoming/outgoing, duration, number

- Total Data
  - Over **400,000 hours continuous human behavior data** collected over the 2004-2005 academic year.

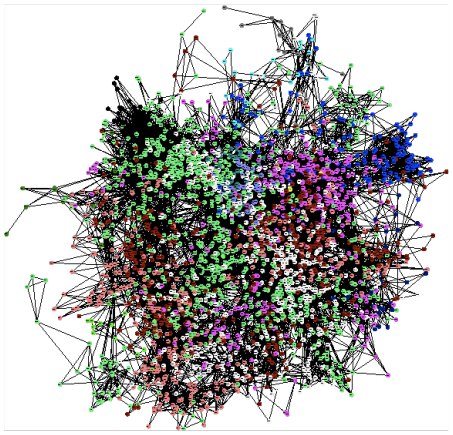
Mika Raento, Antti Oulasvirta, and Nathan Eagle. "Smartphones: an emerging tool for social scientists", *Sociological Methods and Research* (in submission).

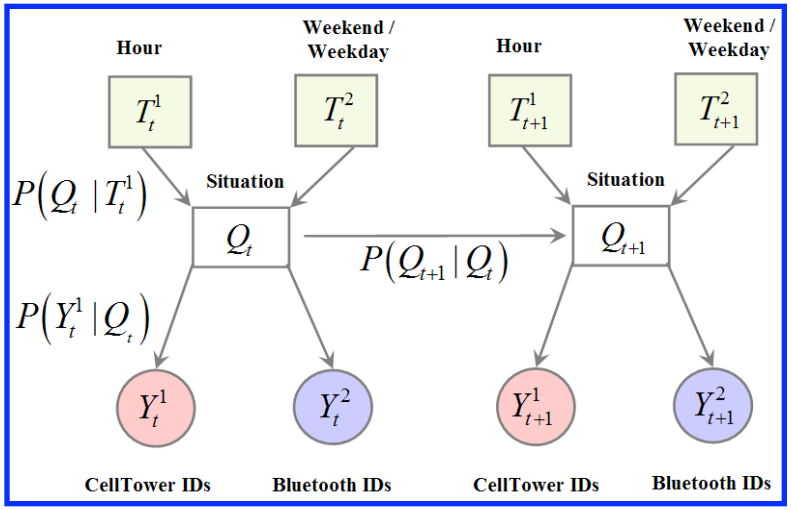


N = 1 HUNDRED

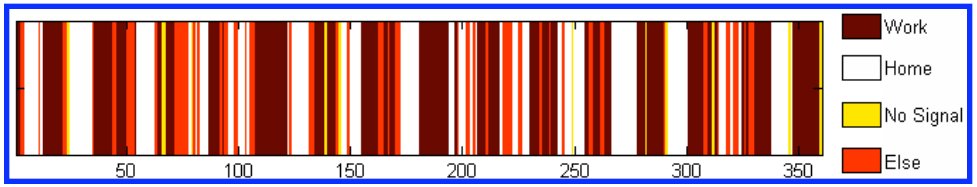
High-Level Situation Classification

- Probabilistic Graphical Models for Data Filtering
  - Conditioned HMM

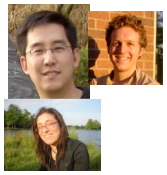




$$Q \in \{Home, Work, Elsewhere, No\ Signal\}$$



Eagle, N., Quinn, J., Clauset, A., "Methodologies for Continuous Cellular Tower Data Analysis", to appear Pervasive'09.



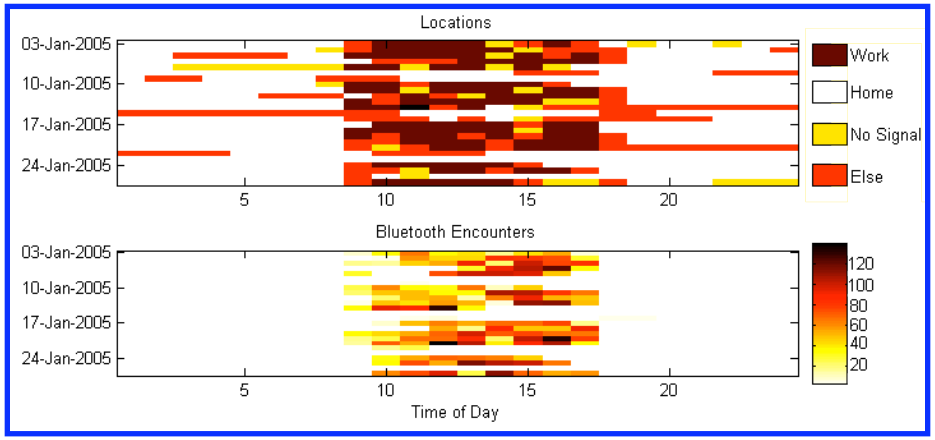
N = 1 HUNDRED

The Entropy of Life

- Shannon Information Entropy Applied to Everyday Life
  - Estimate of the amount of structure / randomness in a subject's routine.

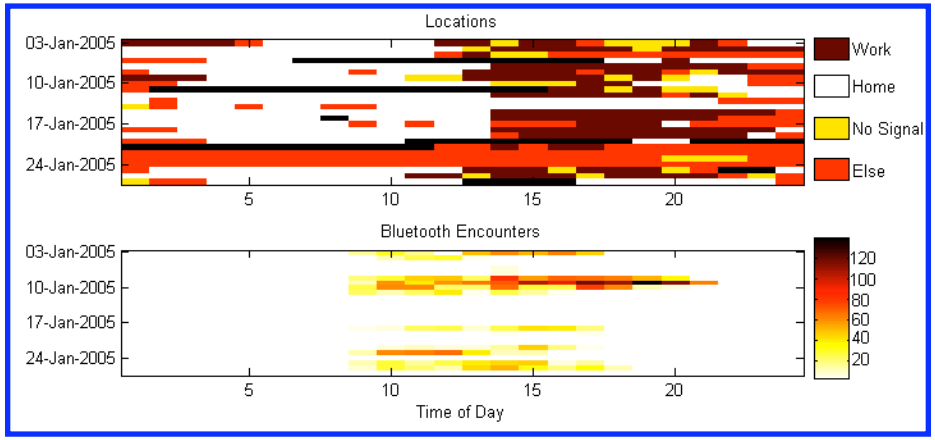
$$H(I) = -\sum_{j=1}^{n} p(j) \log_2 p(j)$$

Low Entropy Subject, I_1



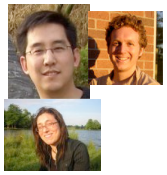
$H(I_1) = 30.9$

High Entropy Subject, I_2



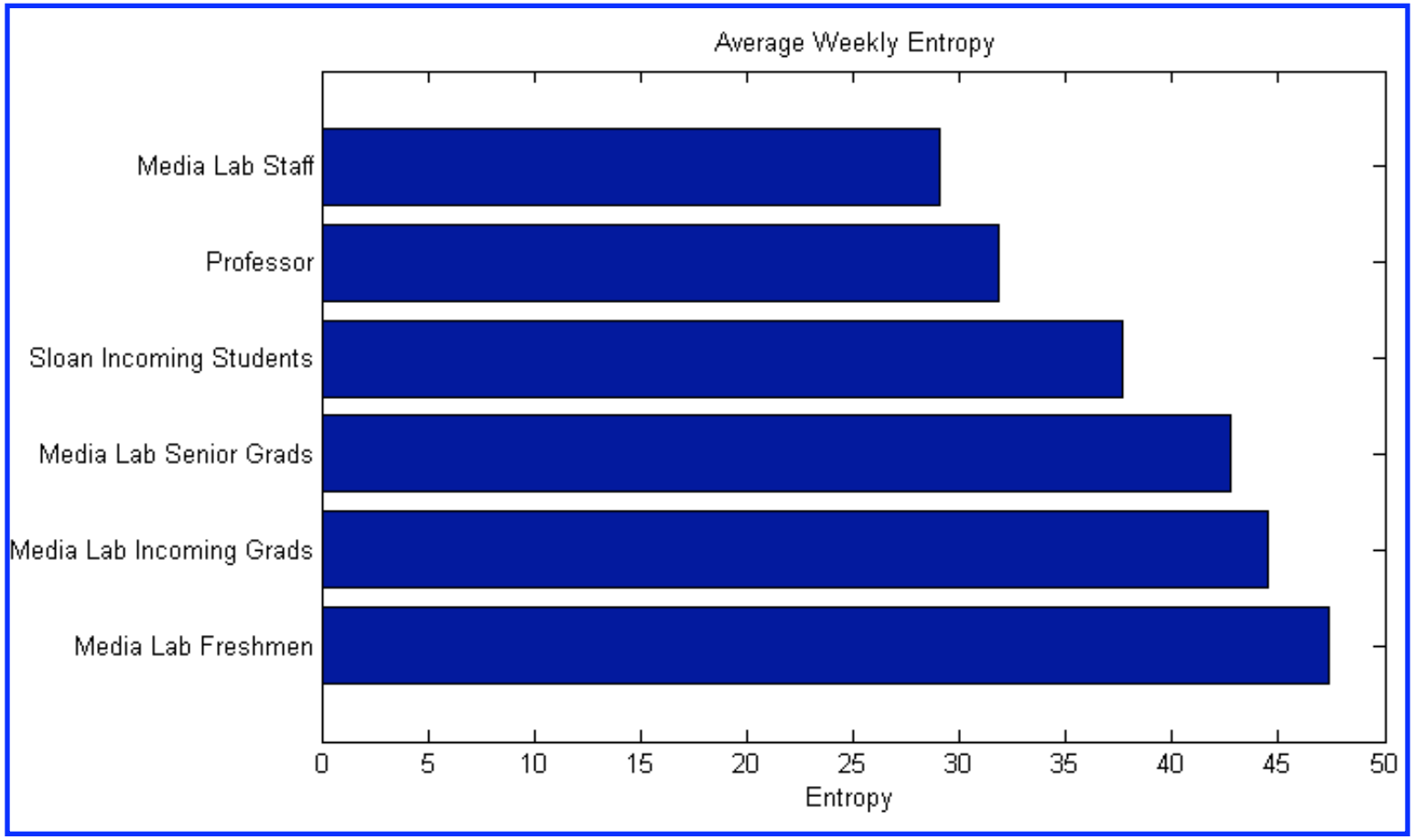
$H(I_2) = 48.5$

N. Eagle and A. Pentland (2006), "Reality Mining: Sensing Complex Social Systems", Personal and Ubiquitous Computing, Vol 10, #4, 255-268.

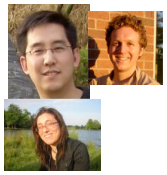


N = 1 HUNDRED

Behavioral Entropy

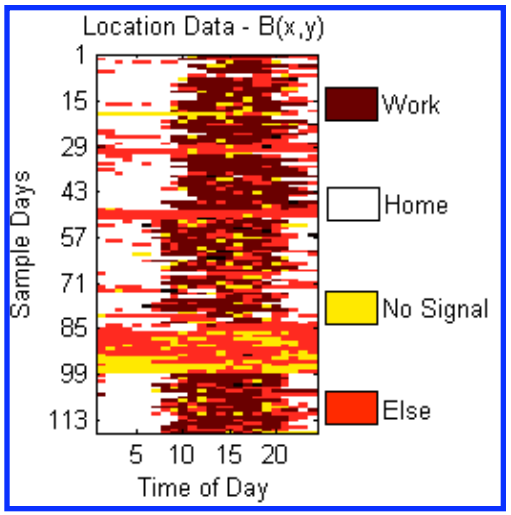


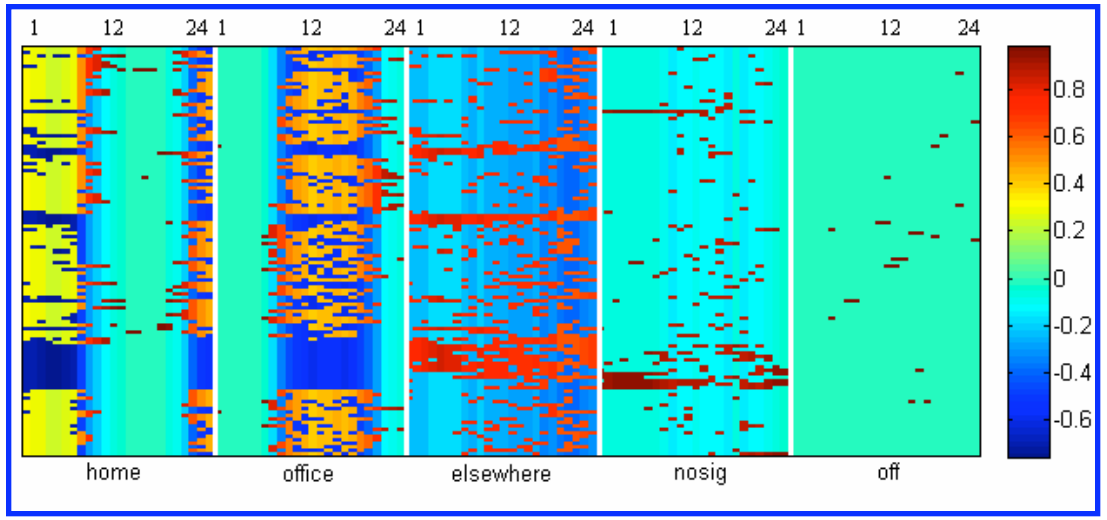
Which Demographic is most Infectious?



N = 1 HUNDRED

Eigenbehaviors: Transformation





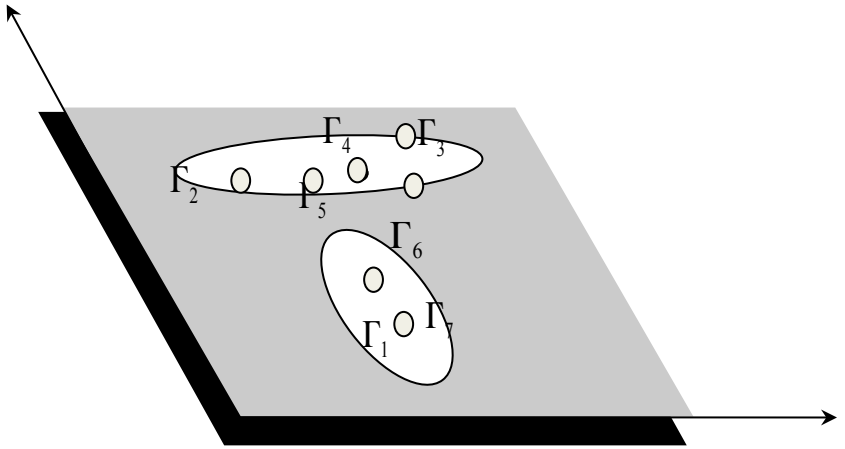
$$\Psi = \frac{1}{N}\sum_{i=1}^{N}\Gamma_i$$

$$\Phi_i = \Gamma_i - \Psi$$

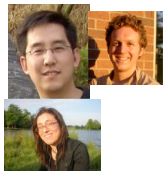
$$A = \left[\Phi_1;\Phi_2;\Phi_3;\ldots\Phi_N\right]$$

$$\lambda_k = \frac{1}{N}\sum_{i=1}^{N}\left(u_k\Phi_i\right)^2$$

Individual Behavior Space

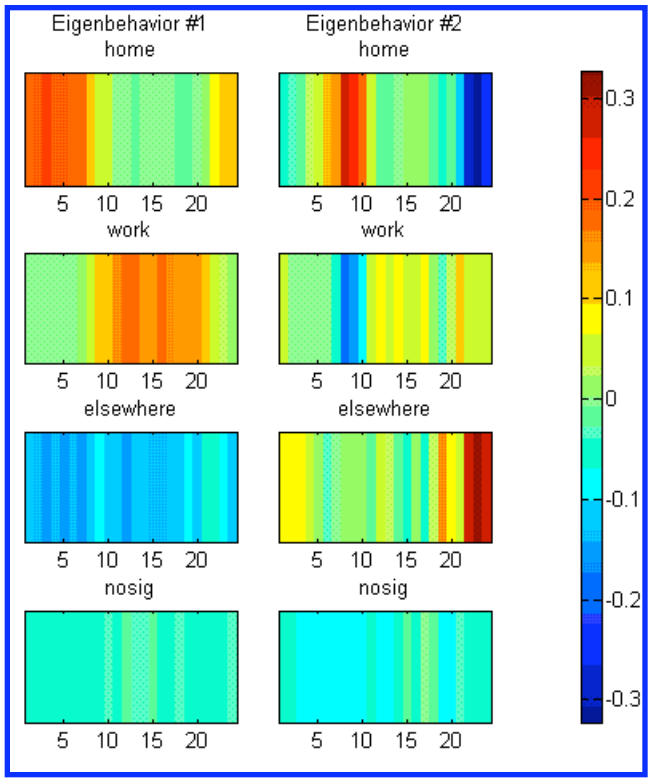


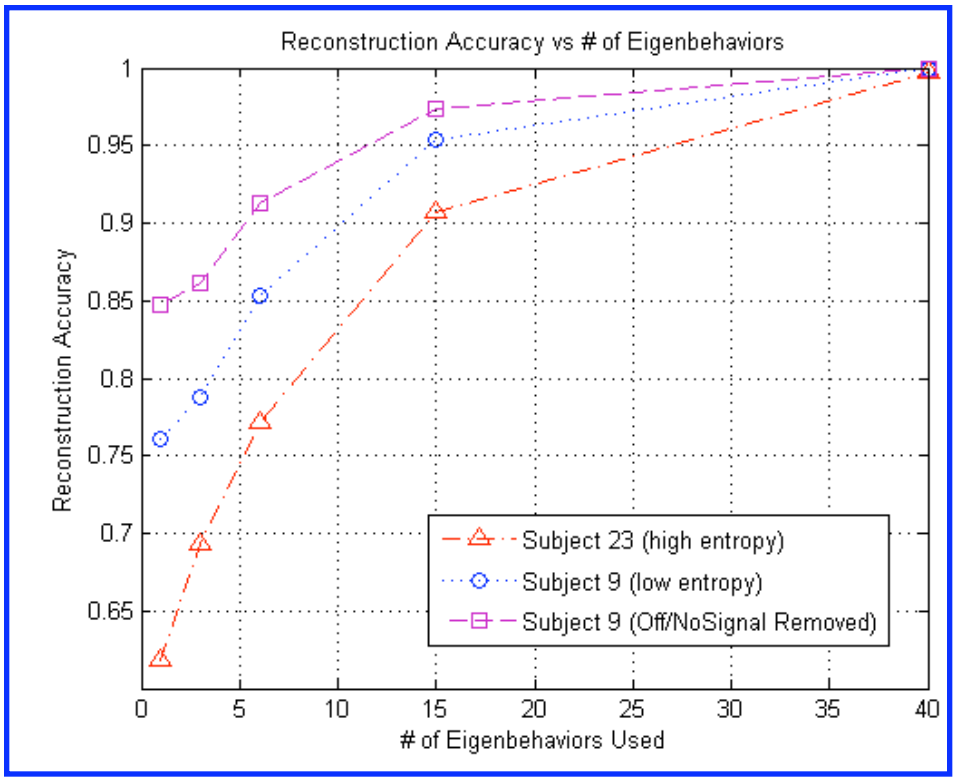
Turk, M., and Pentland, A., "Eigenfaces for Recognition", *J. of Cognitive Neuroscience*. Vol 3, Number 1., (1991) 71-86



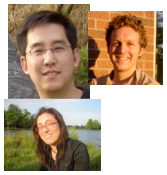
N = 1 HUNDRED

Eigenbehaviors: Behavior Space





Nathan Eagle and Alex Pentland. (2008) "Eigenbehaviors: Identifying Structure in Routine", *Behavioral Ecology and Sociobiology.* (in press)



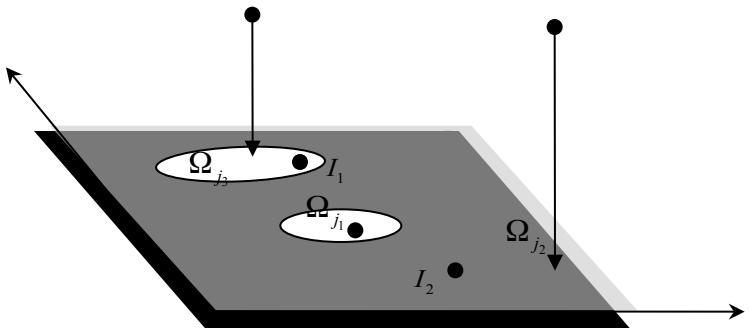
N = 1 HUNDRED

Eigenbehaviors: Affiliation Inference

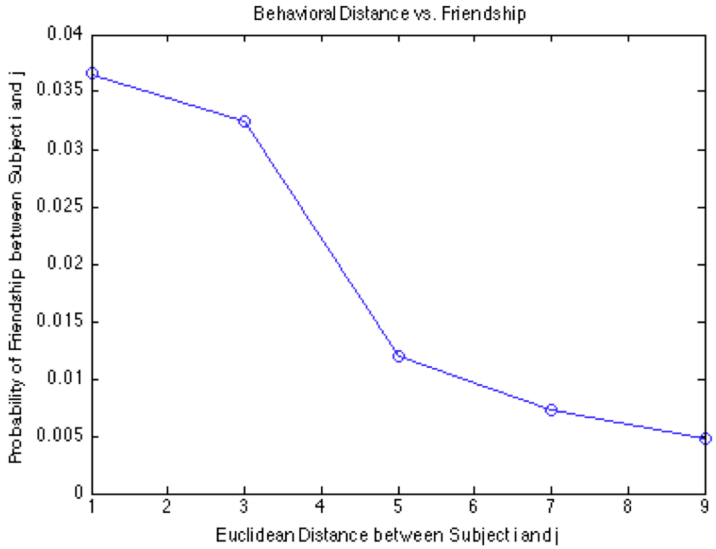
$$\Phi^j = I - \Psi^j$$

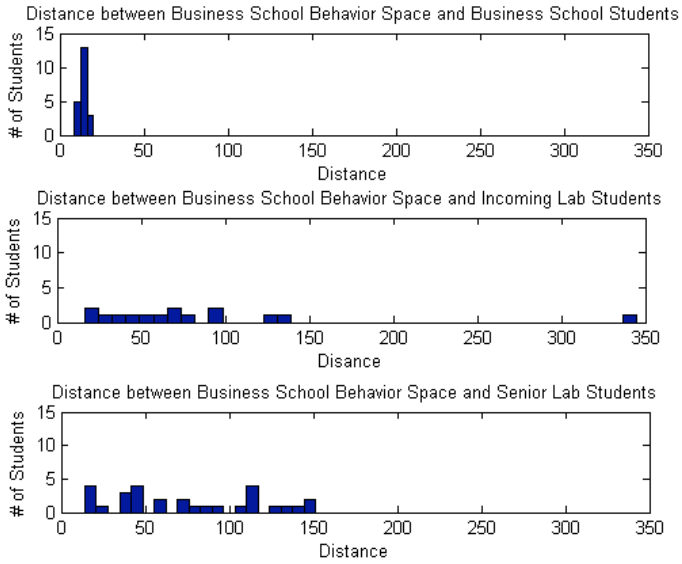
$$\Phi_b^j = \sum_{i=1}^{M'} \omega_i^j u_i^j$$

$$\varepsilon_j^2 = \left\| \Phi^j - \Phi_b^j \right\|^2$$



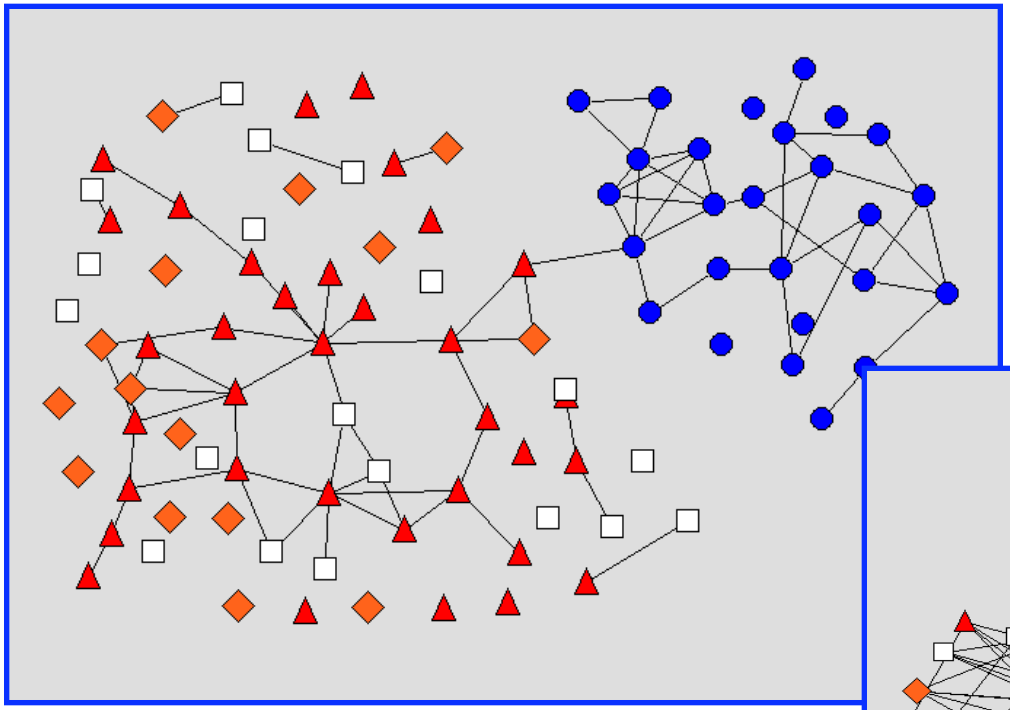
Group *j* Behavior Space



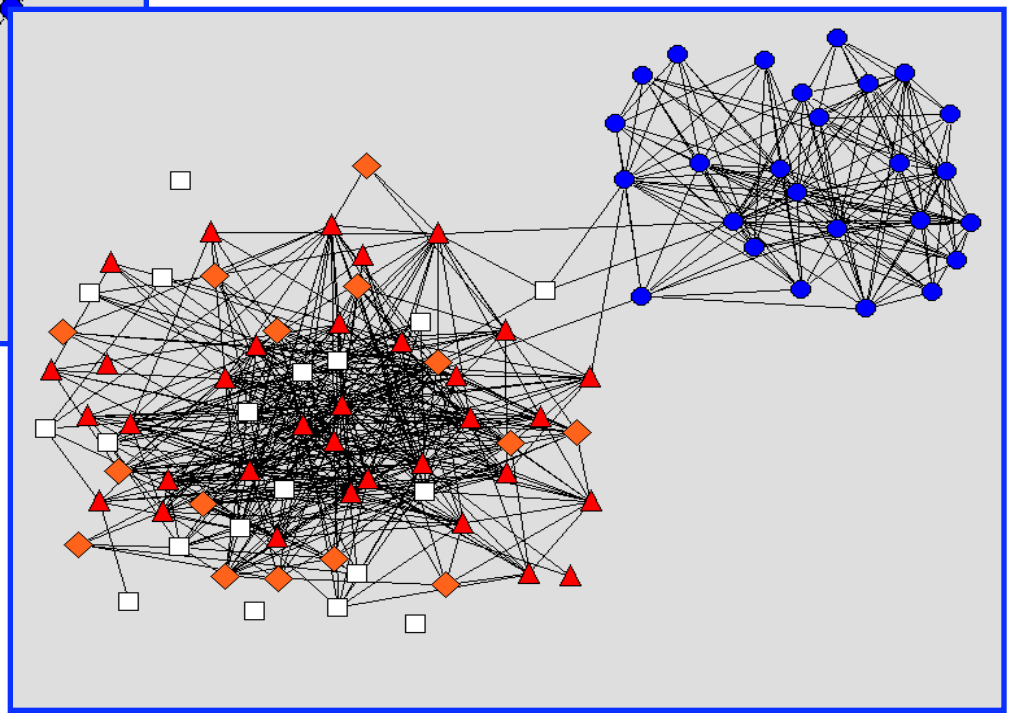


Nathan Eagle and Alex Pentland. (2008) "Eigenbehaviors: Identifying Structure in Routine", *Behavioral Ecology and Sociobiology.* (in press)

Friendship vs. Proximity Networks



Self-Report Friendship



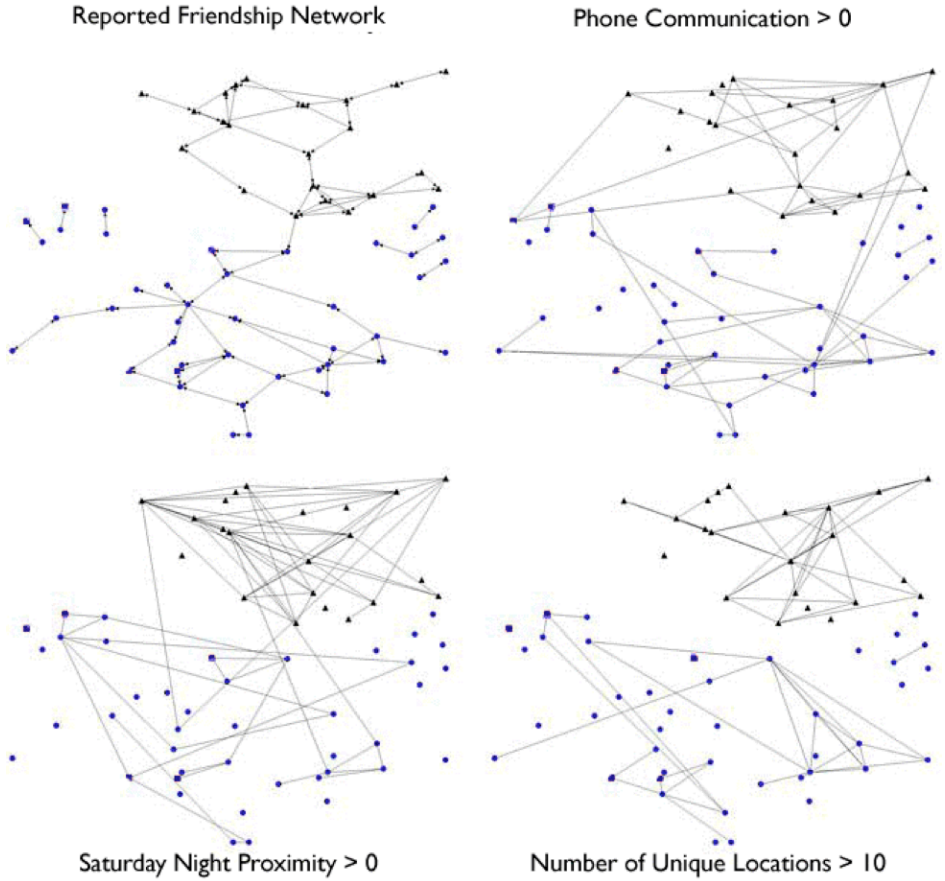
1-Day Proximity

N = 1 HUNDRED

MIT: 63 relationships, 1.3M proximity edges, T = 9 months

- Dyadic Variables for Friendship Inference

	Friends		Not Friends	
	avg	std	avg	std
Total Proximity (minutes / day)	**72**	150	**9.5**	36
Saturday Night Proximity (minutes / week)	**7.3**	18	**.20**	1.7
Proximity with no Signal (minutes / day)	**12**	20	**2.9**	20
Total Number of Towers Together	**20**	36	**3.5**	4.4
Proximity at Home (minutes / day)	**3.7**	8.4	**.32**	2.2
Phone Calls / day	**.11**	.27	**.001**	.017

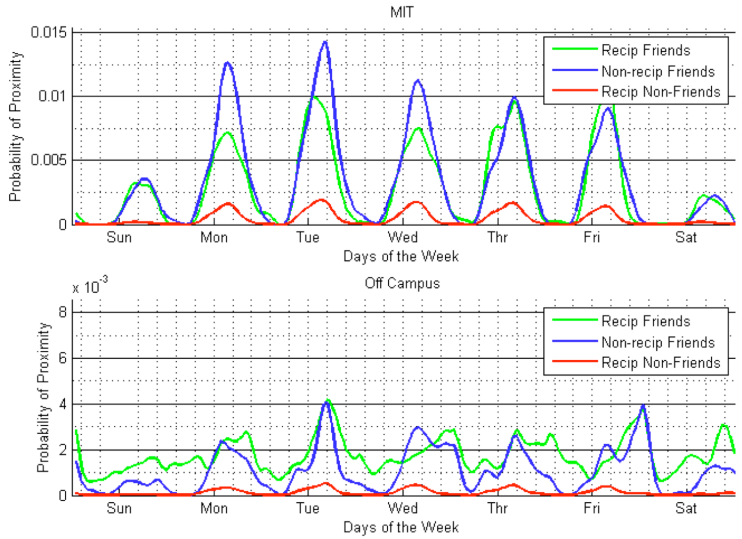


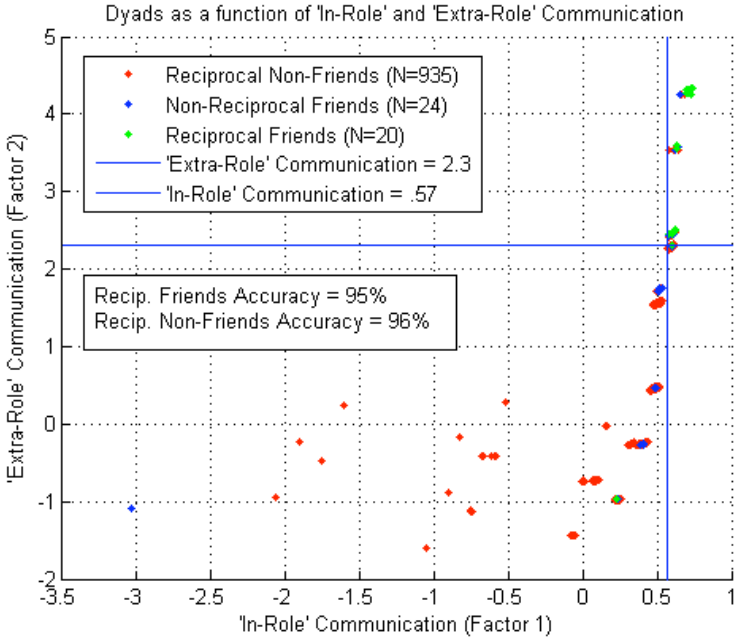
COLLABORATORS: David Lazer (Harvard), Sandy Pentland (MIT)

Eagle, N., Pentland, A., Lazer, D. "Inferring Social Network Topology", *PNAS*. (in press).

N = 1 HUNDRED

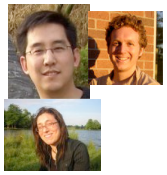
MIT: 63 relationships, 1.3M proximity edges, T = 9 months





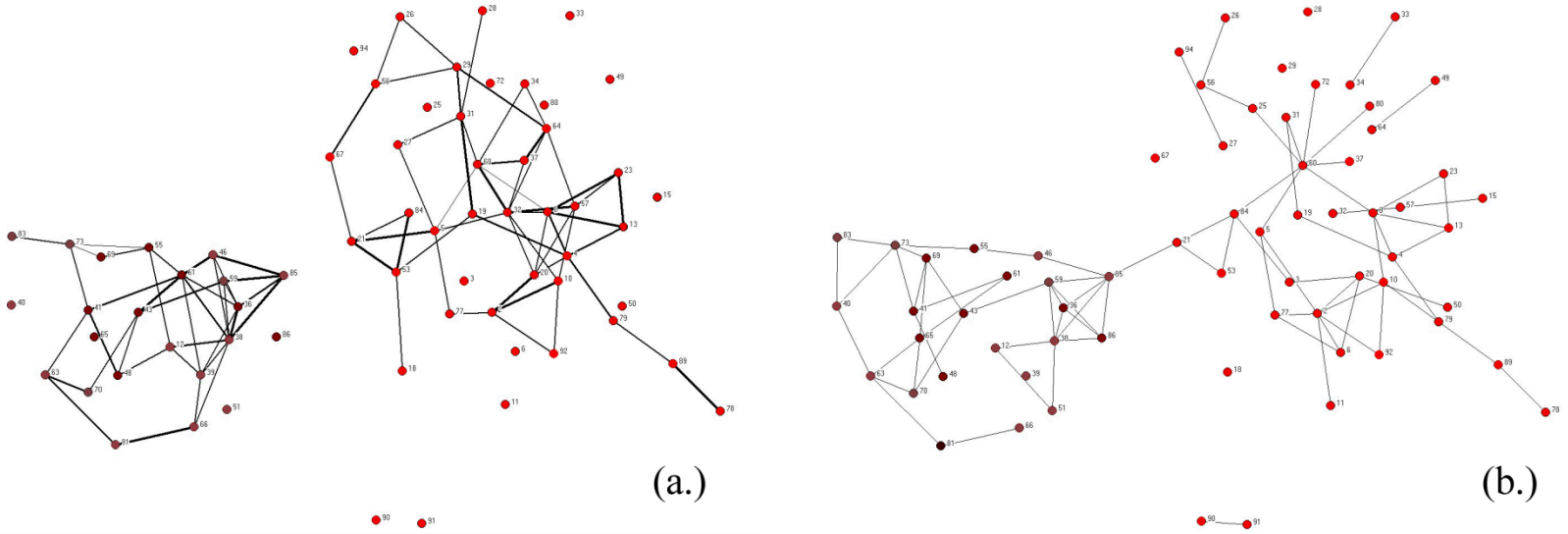
COLLABORATORS: David Lazer (Harvard), Sandy Pentland (MIT)

Eagle, N., Pentland, A., Lazer, D. "Inferring Social Network Topology", *PNAS*. (in press).

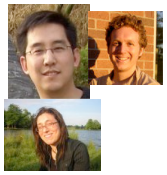


N = 1 HUNDRED

Inferred v. Reported Network



Nathan Eagle, Alex Pentland, and David Lazer. “Inferring Social Network Structure using Mobile Phone Data”, *PNAS* (in submission).

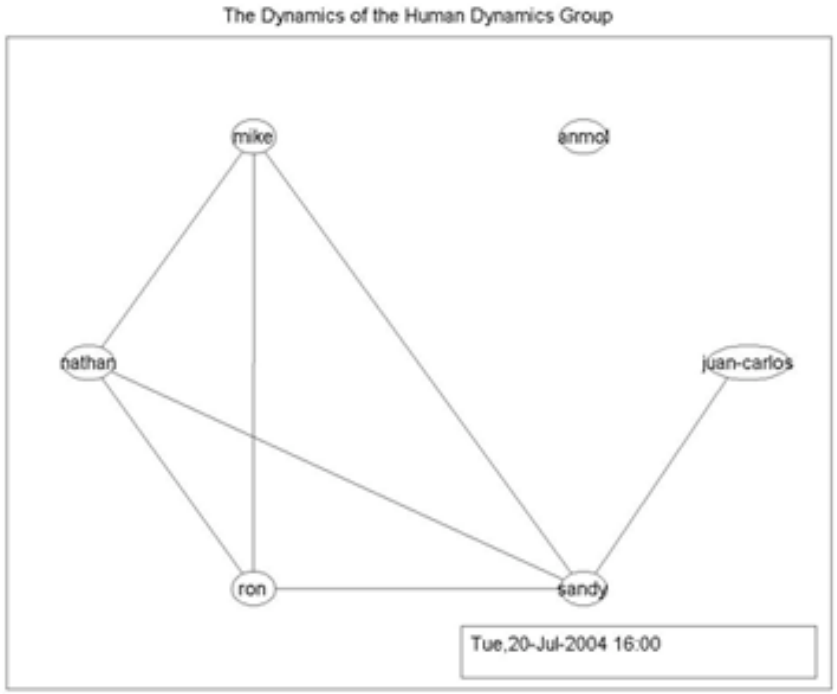


N = 1 HUNDRED

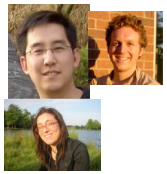
Turning Real Data into Something we're good at

Series of Static Snapshots = $G_t=\{N_t;E_t\}$

$$B_{i,j}^{(t)} = \begin{cases} 1 & \text{if vertices } i \text{ and } j \text{ are ever connected} \\ & \text{between time } t \text{ and } t+\Delta, \\ 0 & \text{otherwise.} \end{cases}$$

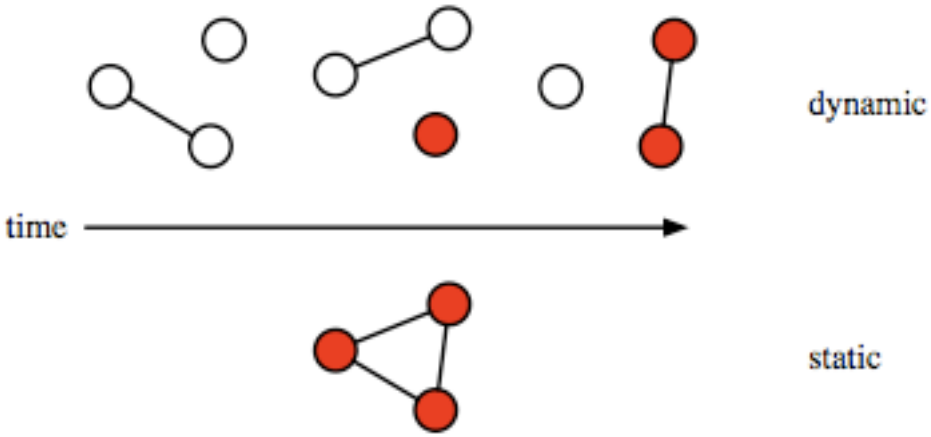


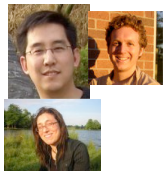
Δ = 1 hour



N = 1 HUNDRED

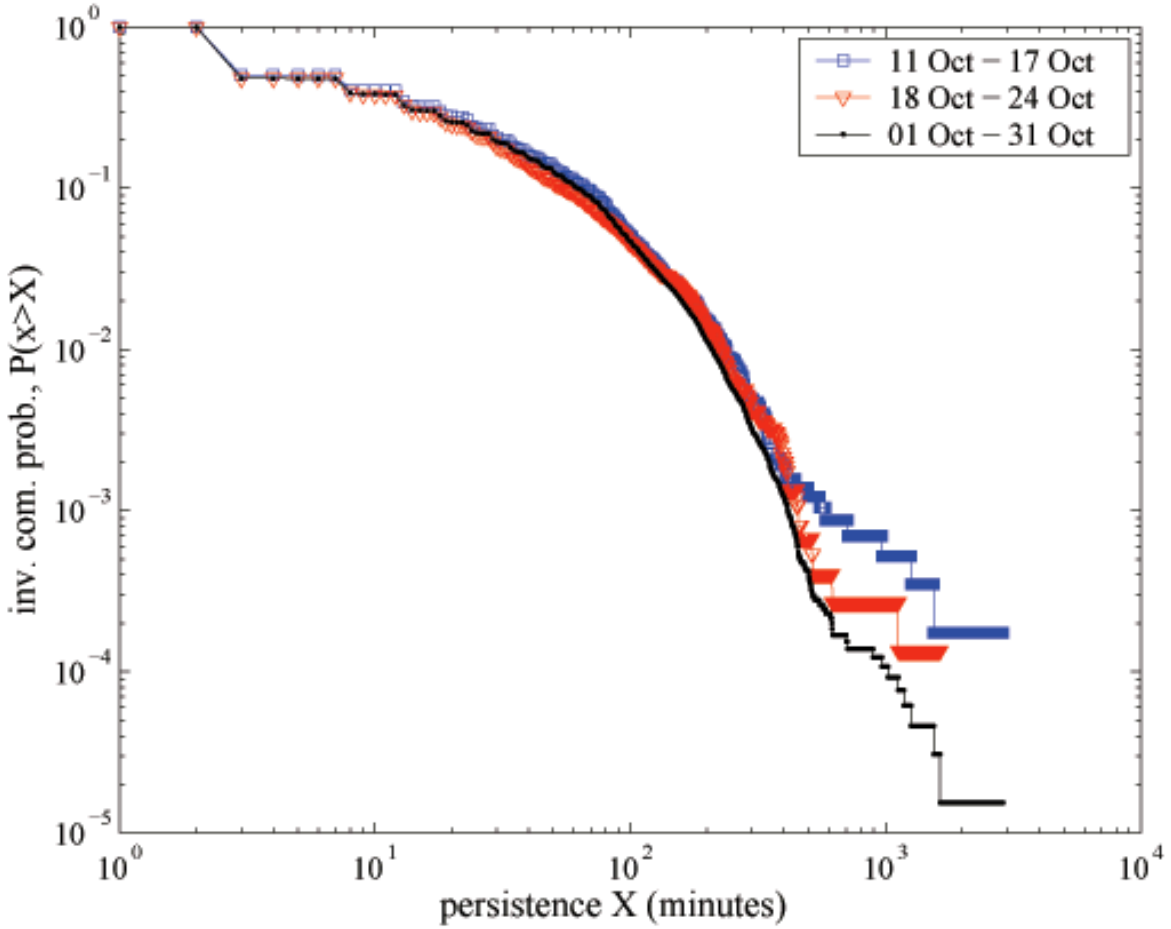
Misdetection of Contagion Spread

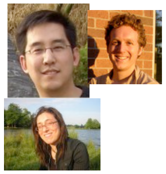




N = 1 HUNDRED

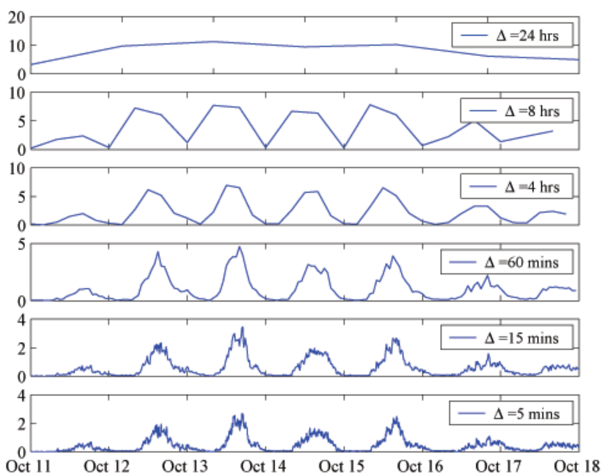
Edge Persistence





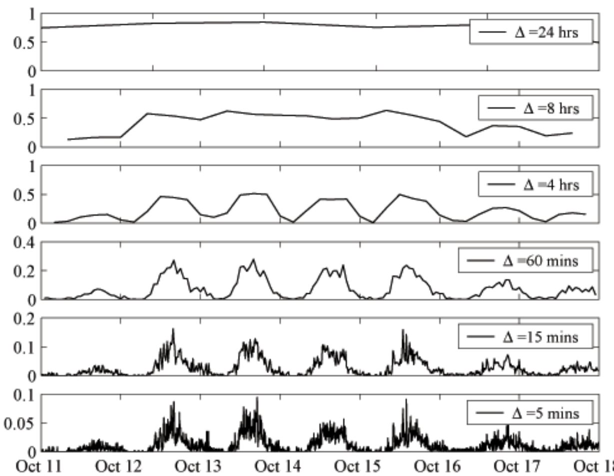
N = 1 HUNDRED

Washing out Dynamics via Sampling



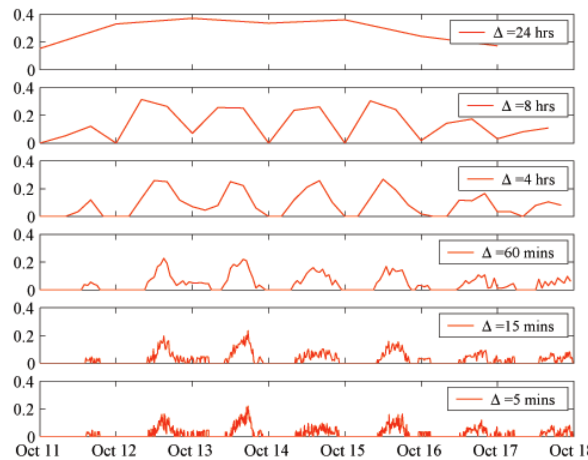
Mean Degree, $\overline{k}$

$$\overline{k} = \frac{k_{total}}{n}$$



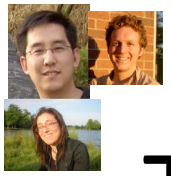
Clustering Coefficient, C

$$C = \frac{1}{n}\sum_{i=1}^{n} \frac{(\text{number of triangles centered on vertex } i)}{(\text{number of triples centered on vertex } i)}$$



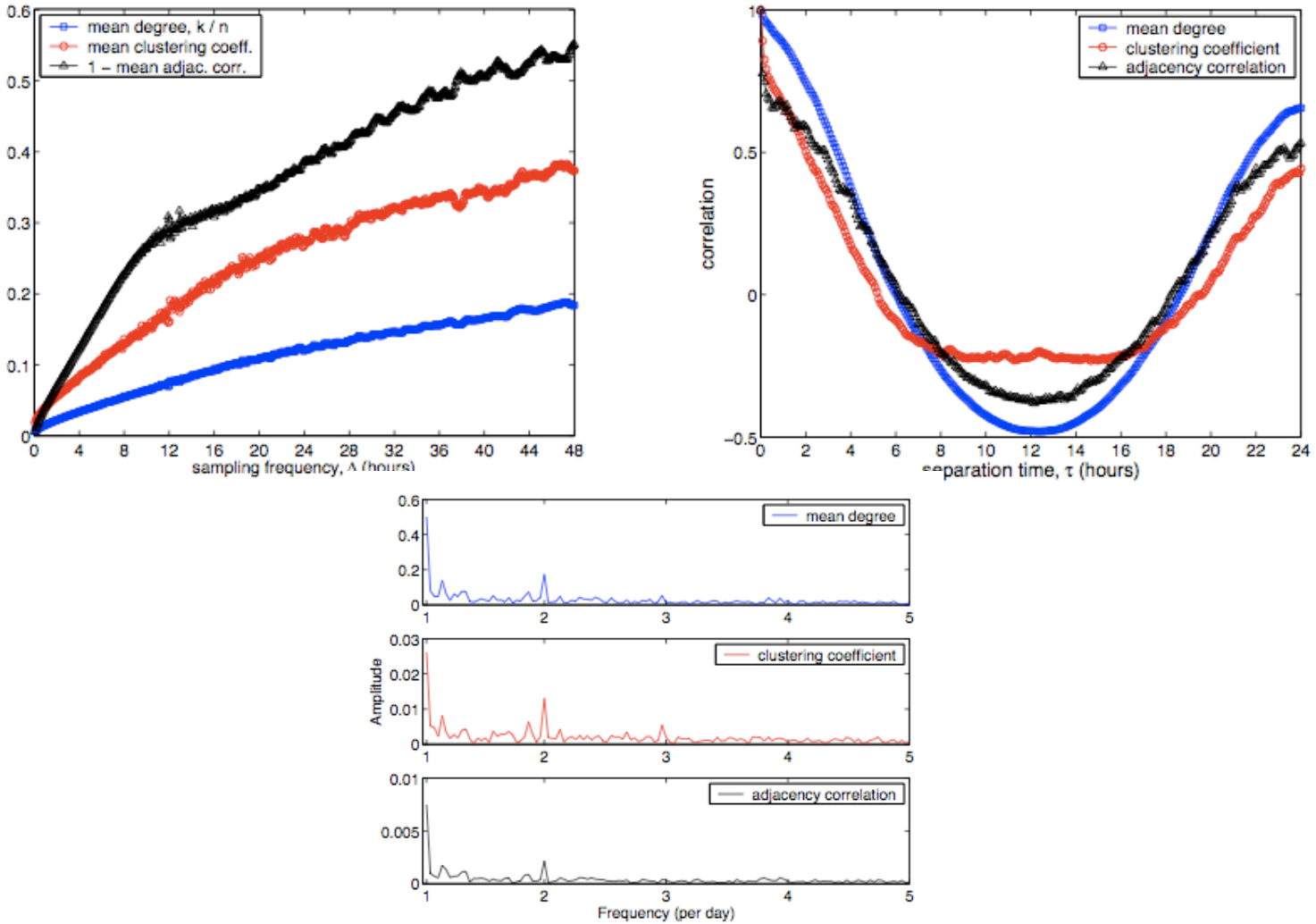
Network Correlation, $1-\gamma$

$$\gamma_j = \frac{\sum_{i \in N(j)} A_{i,j}^{(t_1)} A_{i,j}^{(t_2)}}{\sqrt{\left(\sum_{i \in N(j)} A_{i,j}^{(t_1)}\right)\left(\sum_{i \in N(j)} A_{i,j}^{(t_2)}\right)}}$$

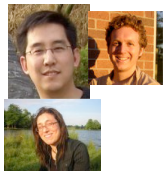


N = 1 HUNDRED

Towards Human Nyquist Sampling?



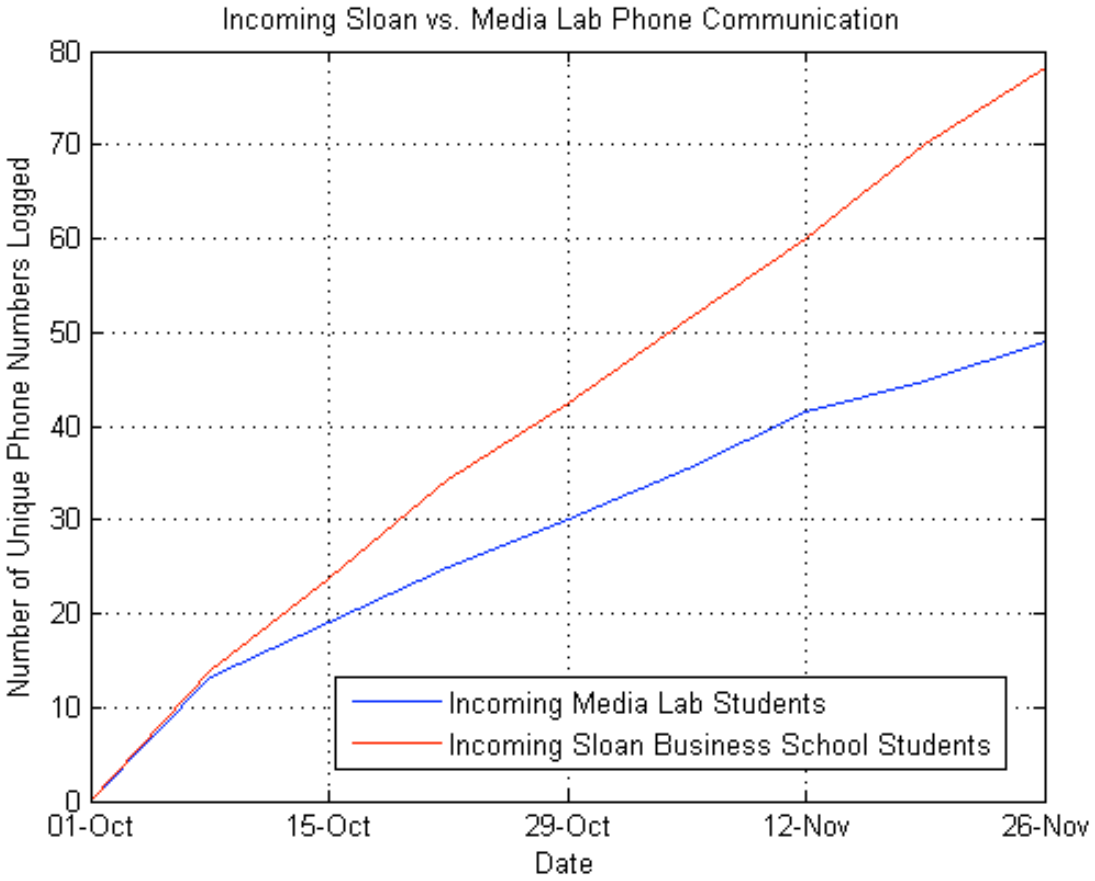
Clauset, A., and Eagle, N., Persistence and periodicity in a dynamic proximity network, DIMACS Workshop on Computational Methods for Dynamic Interaction Networks, 2007.

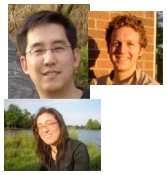


N = 1 HUNDRED

Network Evolution

- Can mobile phone usage reflect an emphasis on ‘networking’ and social network evolution?

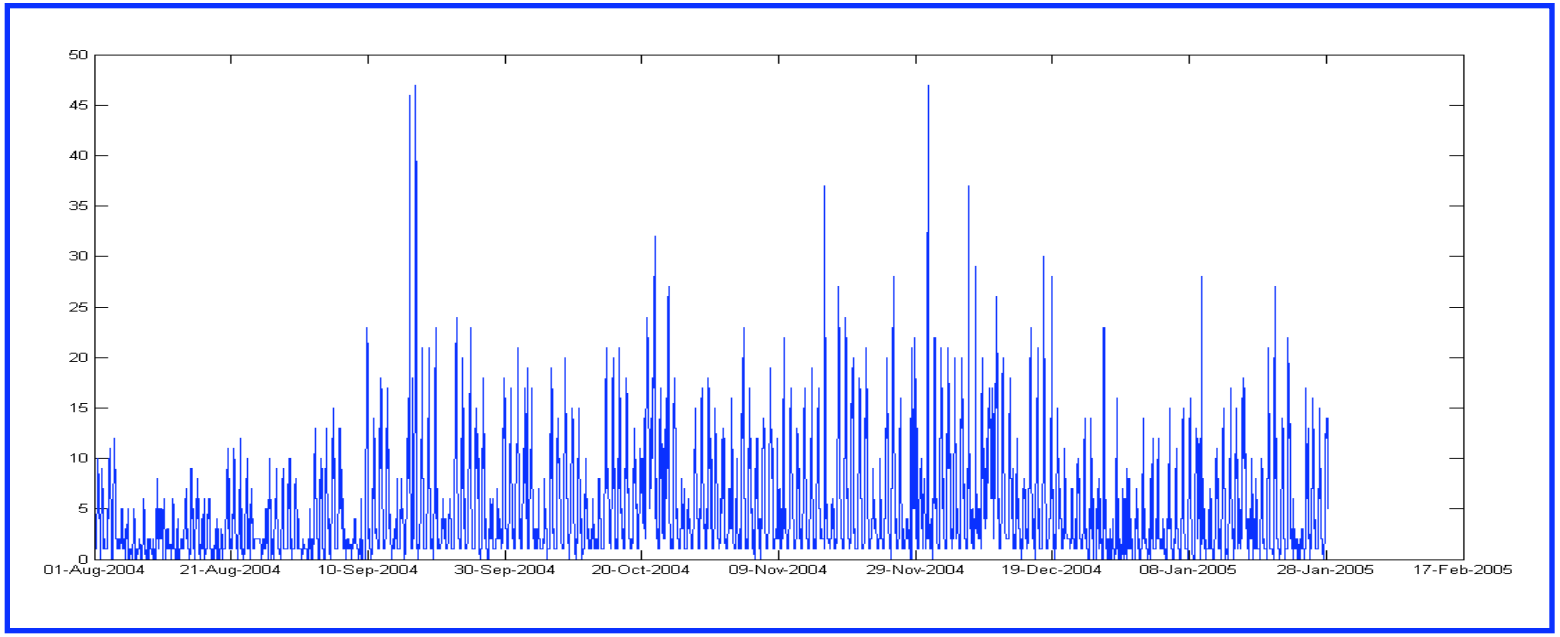


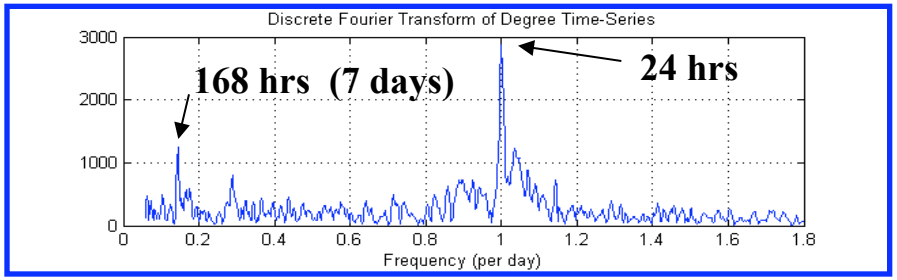


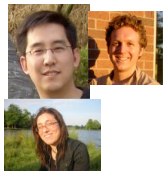
N = 1 HUNDRED

Organizational Rhythms

- How the deadlines of an institution can be seen in the collective behavior of its individual members.



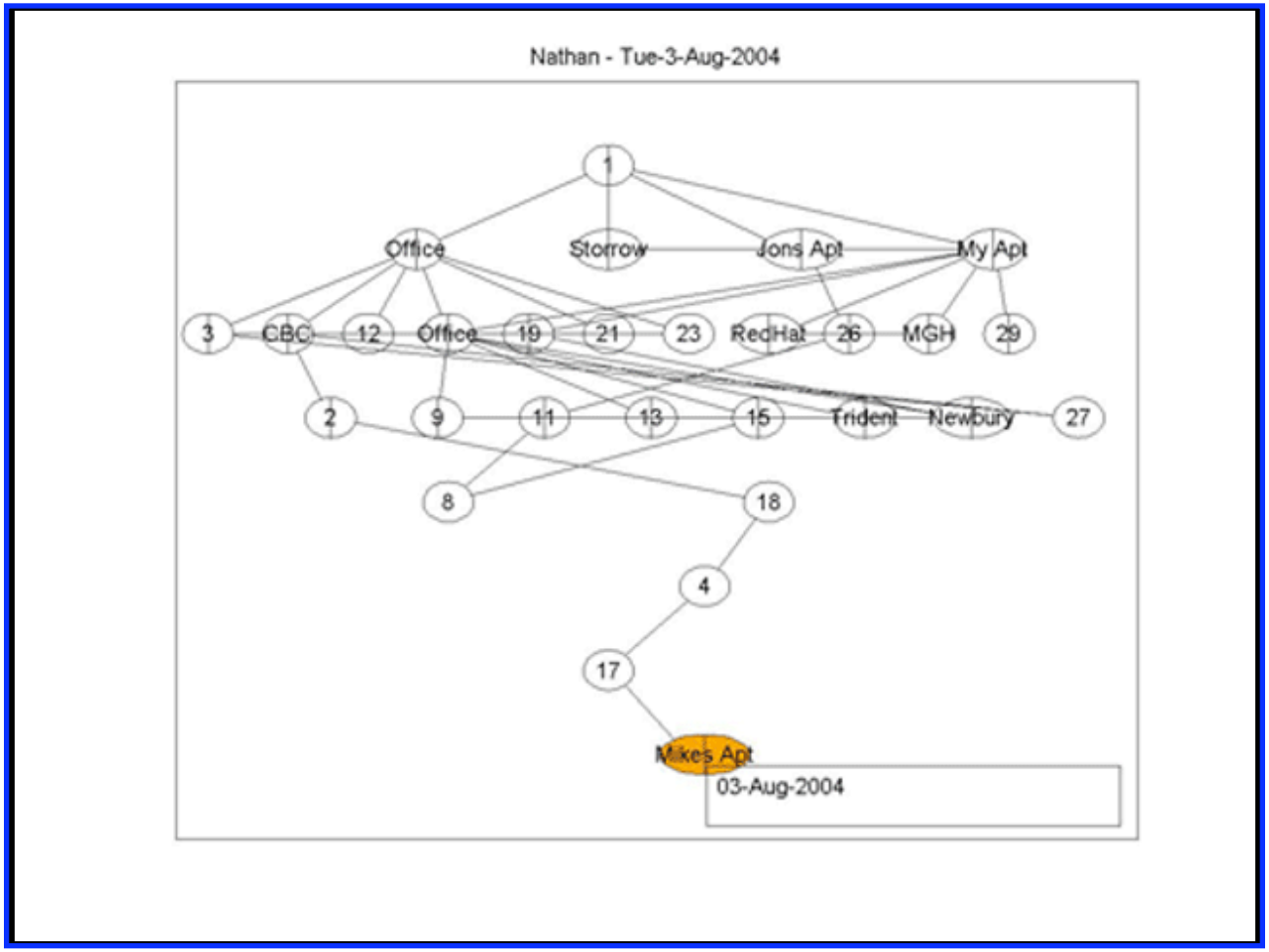


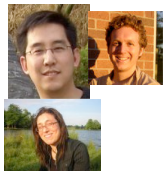


N = 1 HUNDRED

Applications…

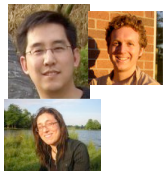
- Automatic Diary Generation:
  - A life log from cell tower IDs




N = 1 HUNDRED

Life Inferences

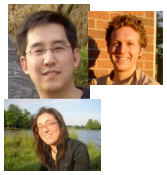
- Class: Sleeping?
  - Location: {Home}
  - Phone status: {idle/charging}
  - Time: {late night / early morning}
  - Alarm Clock: {interval}
- Class: Lunch?
  - Location: {!= office}
  - People: {lunch crowd={Mike, Push, Martin}}
  - Time: {lunchtime}
- Class: Partying?
  - Location: {hang outs ={b-side, sevens, BHP}}
  - People: {party friends={Mike, Jon, Aisling}}
  - Time: {evening / late night}



N = 1 HUNDRED

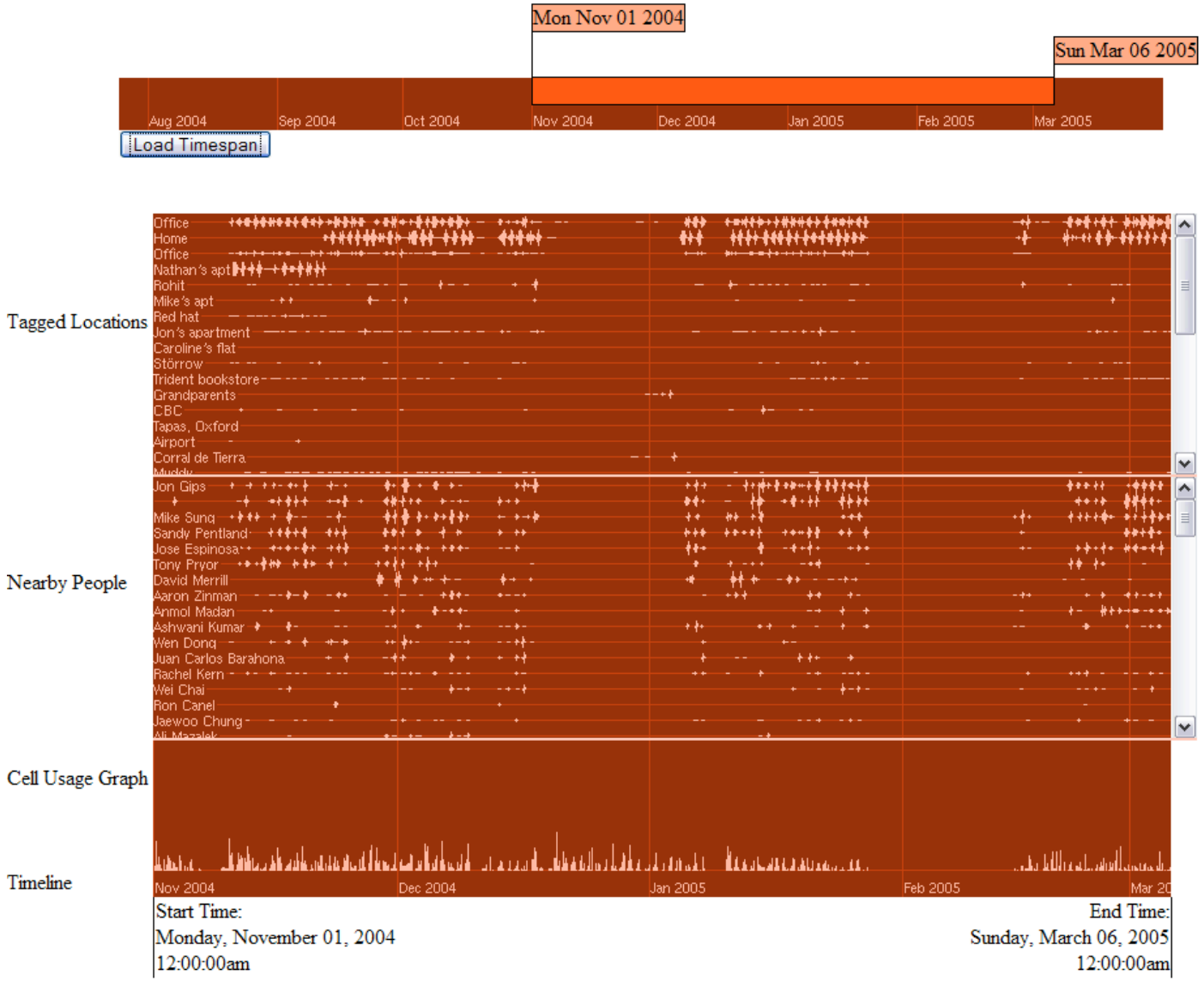
Life Query

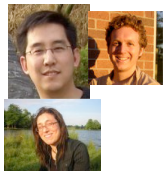
- AutoDiary
  - How much sleep did I get last week?
  - When was the last time I had lunch with Josh?
  - How much time did I spend driving when I was last in Mountain View?
  - Where did I go after leaving Marvin's house last week?
- Prediction
  - What are the chances of seeing Mike in the next hour?
  - How likely is it that Caroline will call me tonight?
  - Will I be in lab this weekend?



N = 1 HUNDRED

Automatic Diary

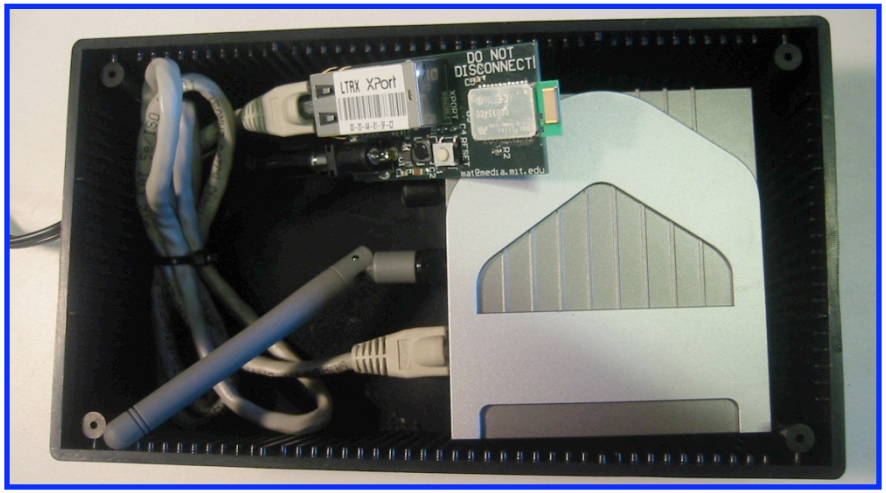




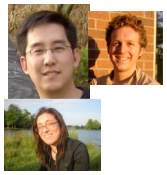
N = 1 HUNDRED

BlueDar : Bluetooth Radar

- Currently Deployed around MIT
  - Infinite Corridor, Media Lab, Muddy Charles Pub, Sloan Business School, Student Center, …
- Coming Soon...
  - Cafeterias
  - Elevators
  - Gym
  - …

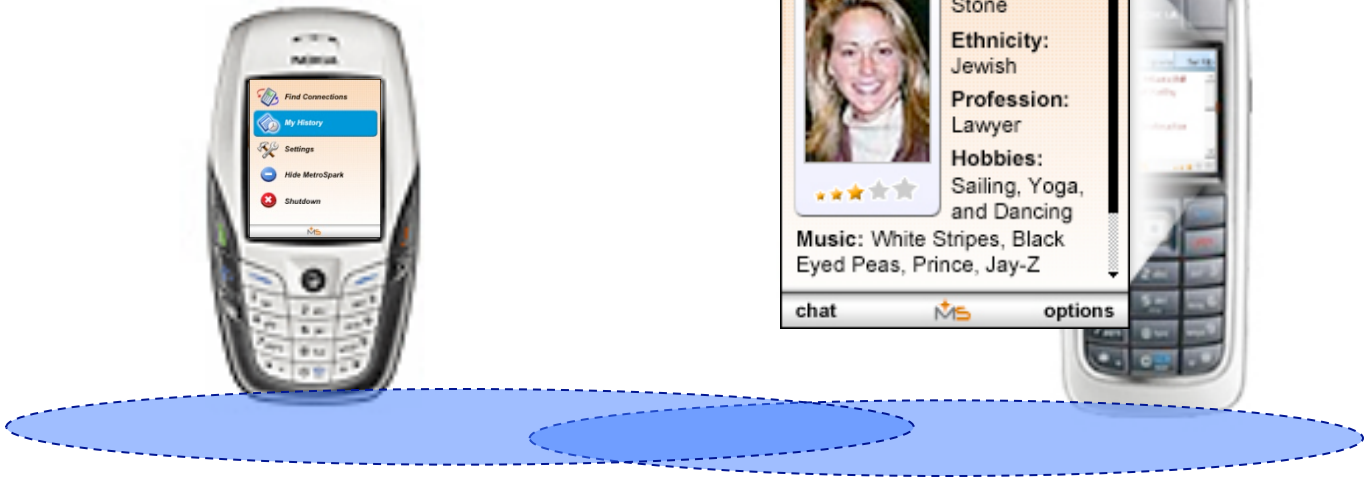


Nathan Eagle, "Can Serendipity Be Planned?", *MIT Sloan Management Review*, Vol. 46, No. 1, pp 10-14, 2004.



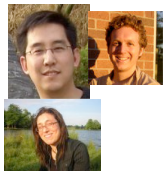
N = 1 HUNDRED

MetroSpark



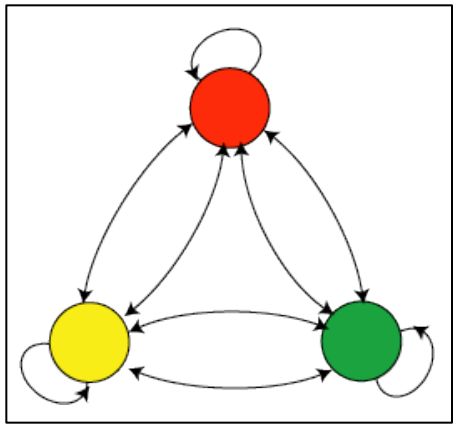
Nathan Eagle and A. Pentland, “Mobile Matchmaking: Proximity Sensing and Cuing”, *IEEE Pervasive Computing*, 4 (2): 28-34, 2005.

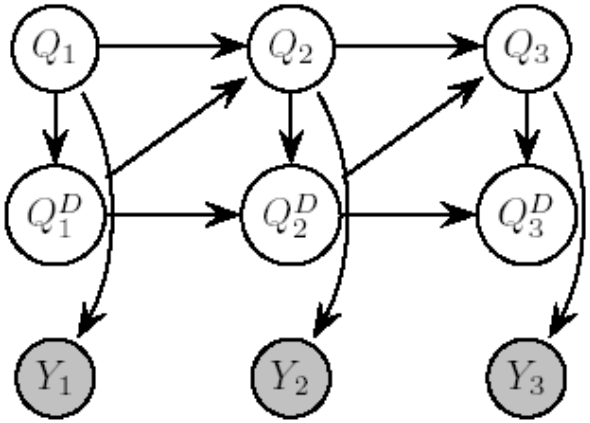
Nathan Eagle and Alex Pentland, "Combined short range radio network and cellular telephone network for interpersonal communications." **U.S. Patent Application Serial No. 60/568,482**. Filed May 6, 2004. MIT ID: 10705T. Assignee: Massachusetts Institute of Technology.

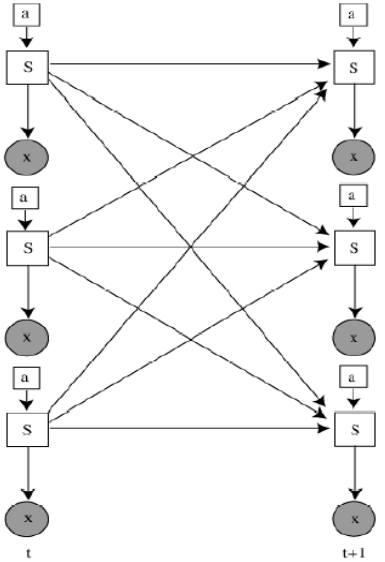


N = 1 HUNDRED

Conversation Modeling

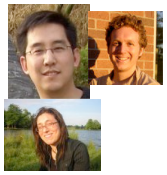






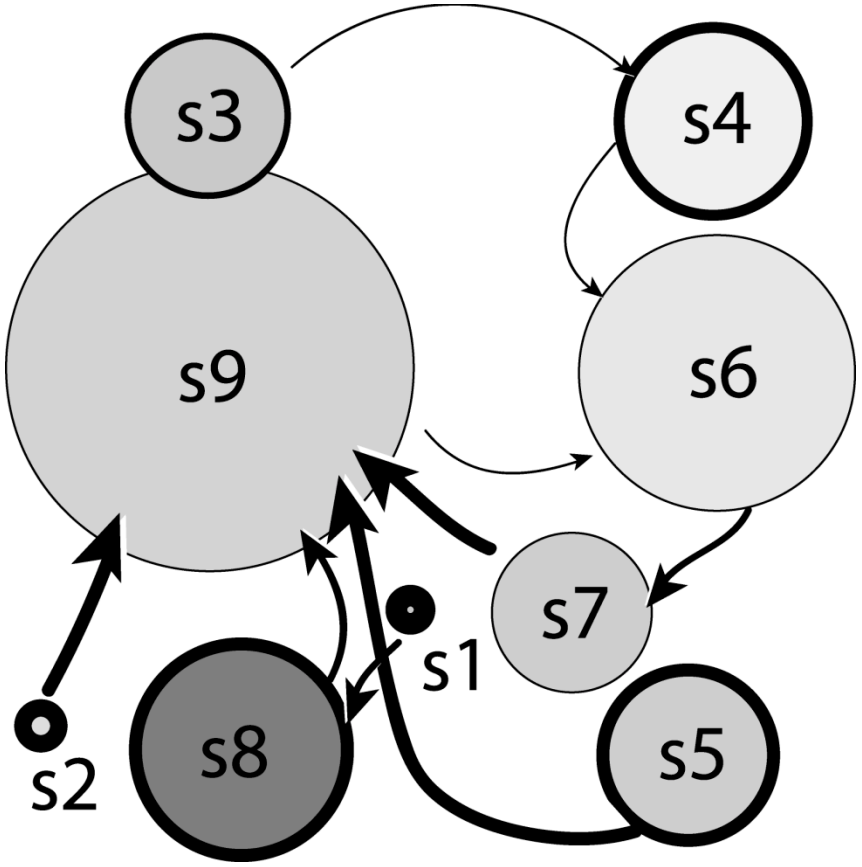
Sumit Basu, Tanzeem Choudhury, Brian Clarkson and Alex Pentland. Learning Human Interactions with the Influence Model. MIT Media Lab Vision and Modeling TR#539, June 2001.

Kevin Murphy. Modeling Sequential Data using Graphical Models. Working Paper, MIT AI Lab, 2002



N = 1 HUNDRED

Prosody Analysis



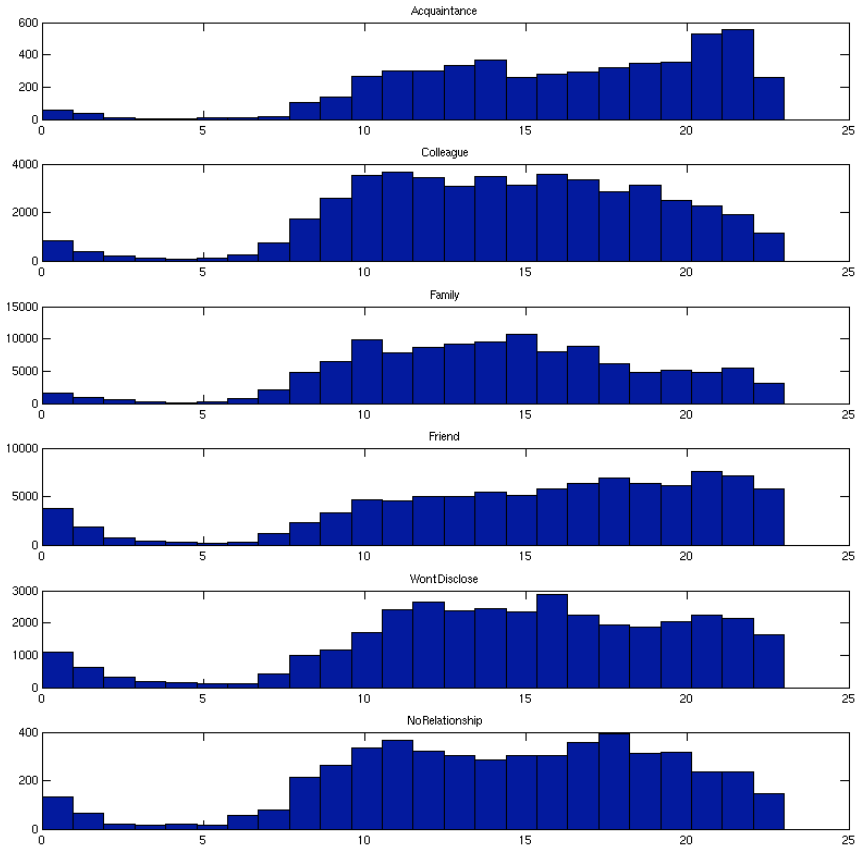
- Speaking Time:
  - *Circle size*
- Transition Probability:
  - *Width of the link*
- Average Interest Level:
  - *Circle color (individual)*
  - *Circle border (group)*

Nathan Eagle and Alex Pentland, "Social Network Computing", Ubicomp 2003: Ubiquitous Computing, *Springer-Verlag Lecture Notes in Computer Science 2864*, pp. 289-296, 2003.

N = 1 THOUSAND

Helsinki: 15k *labeled* relationships,
T = 6 months

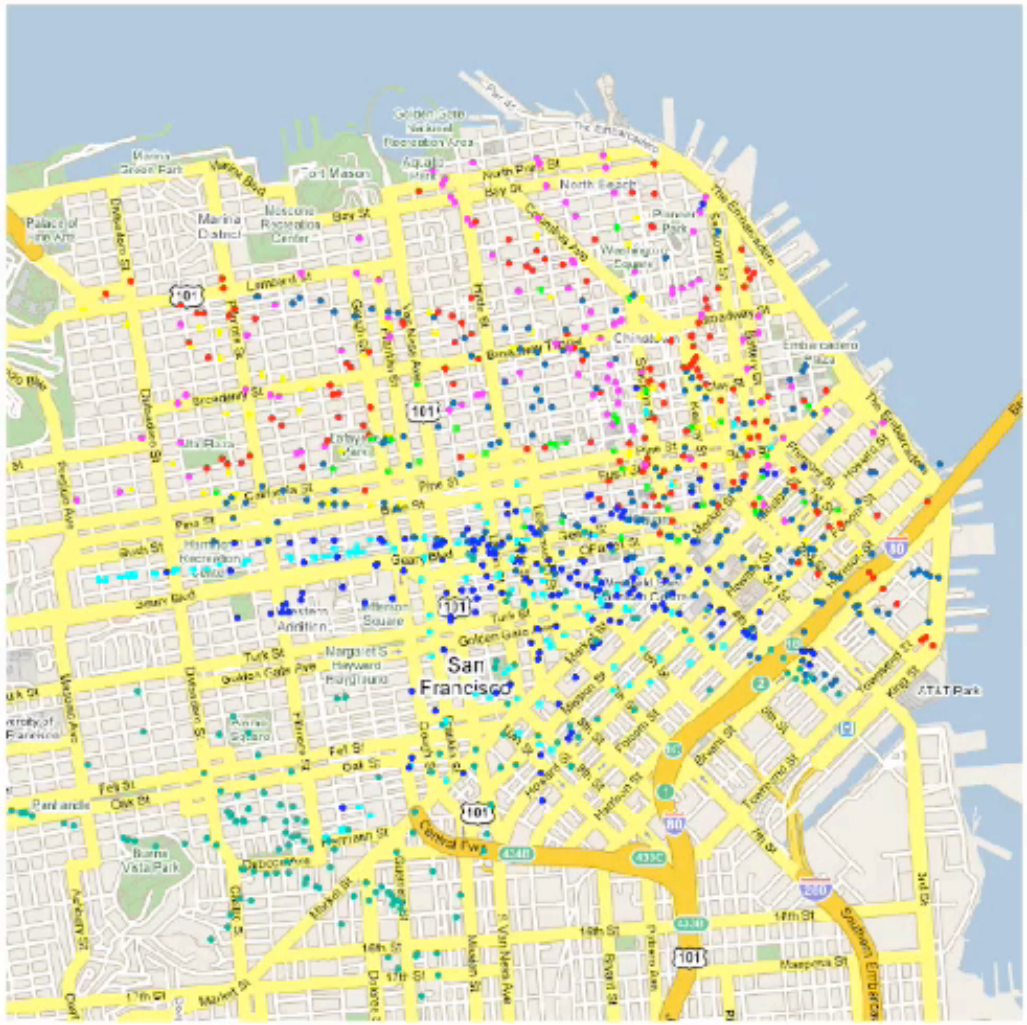
- Do 15,000 labeled, spatial, temporal, contextual edges enable the recognition of generalizable relational signatures?



COLLABORATORS: Hannu Verkasalo (HUT), Cosma Shalizi, Alessandro Rinaldo, Raja Hafiz (CMU)

10 THOUSAND

- Inferring Disease Outbreaks?
  - Ron Hoffeld BACTrack, MIT Lincoln Labs



COURTESY OF SENSENETWORKS: Greg Skibiski, Tony Jebara

N = 10 THOUSAND

US Metropolitan Regions: *Randomly*

Sampled Individuals, T = 1 year

immi Integrated Media Measurement Inc.

IMMI Media Monitoring System

Activity Report - / -

Panelist: 12254
Market: 7 Los Angeles (PST8PDT)

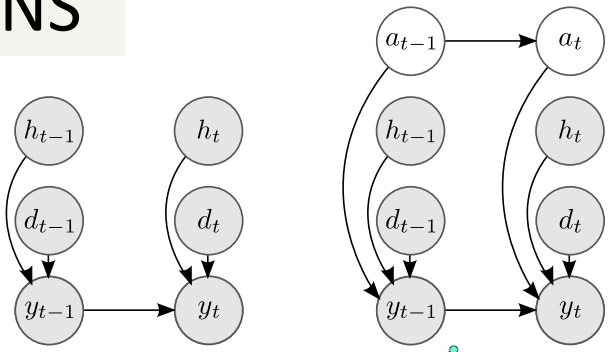
Phone activity detail

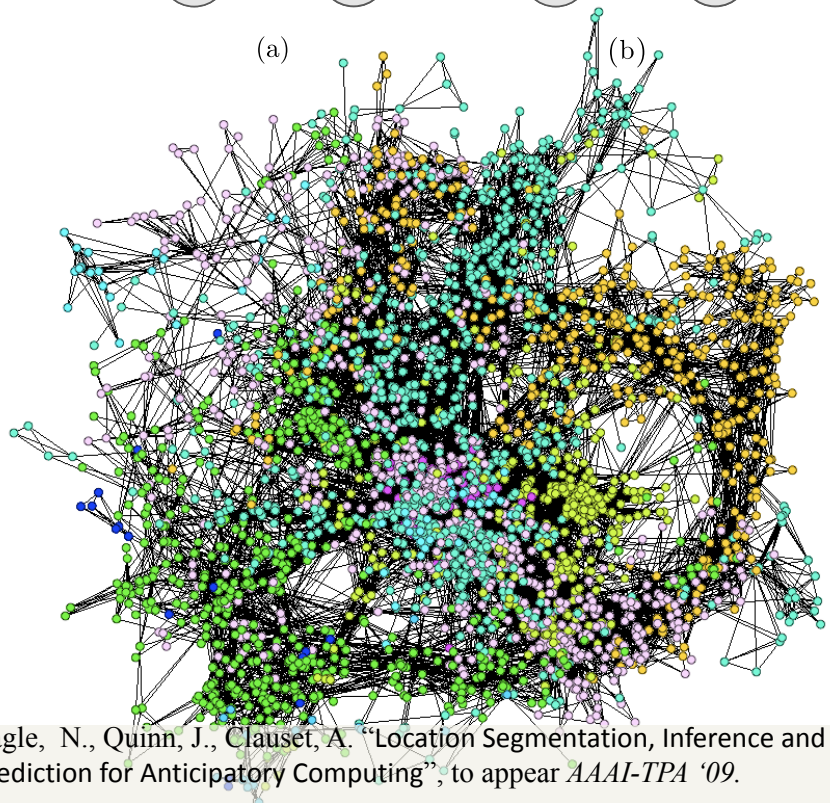
Date	Activity (Packets, Squelch, Sample Err, Samples, Matches, Latency, Battery, Plugged In, Calls)	detail
11/28		
11/29		
11/30		
12/1		

N = 10 THOUSAND

SOME OPEN QUESTIONS

- How to scale DBNs to much higher dimensional state-space?
- Detect Outlier Events?
- Demographic Segmentation?





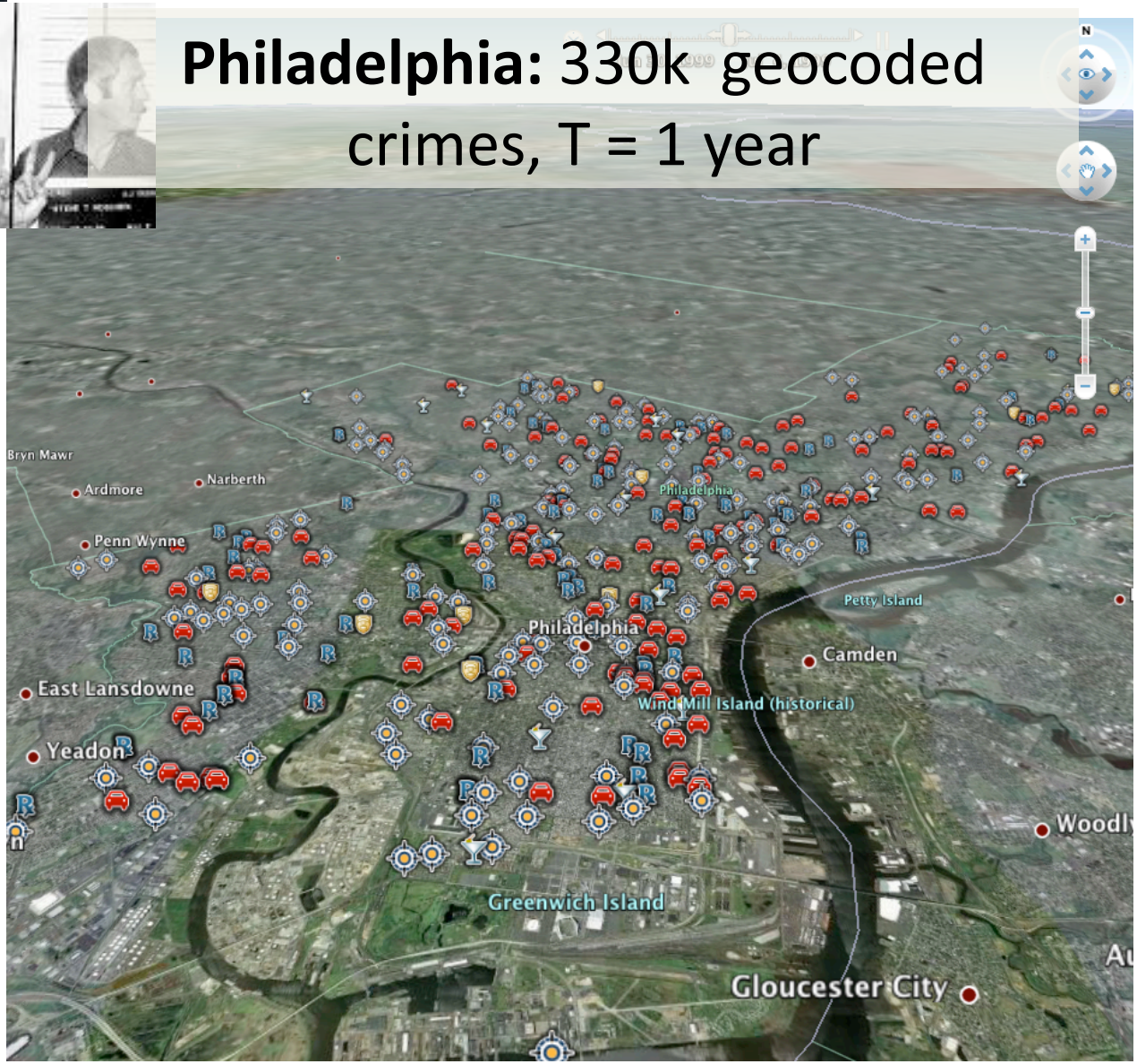
demographic (N)	$\mu_{entropy}$ ($\sigma \times 10^2$)
Age:	
under 35 (107)	30.1 (4.2)
35 and over (108)	28.0 (4.2)
Gender:	
Male (136)	28.3 (4.4)
Female (79)	30.3 (3.8)
Income:	
over $60,000 (73)	34.2 (4.3)
$60,000 and under (140)	26.4 (4.0)
Education:	
College Grad (79)	31.2 (4.3)
No College Degree (125)	27.7 (4.1)

COLLABORATORS: Aaron Clauset and John Quinn

Eagle, N., Quinn, J., Clauset, A. "Location Segmentation, Inference and Prediction for Anticipatory Computing", to appear *AAAI-TPA '09.*

N = 100 THOUSAND

Philadelphia: 330k geocoded crimes, T = 1 year



N = 100 THOUSAND

Philadelphia: 330k geocoded crimes, T = 1 year

- Temporal Dynamics:
  - Does Graffiti lead to Homicide?
- Contagion = Crime "Waves"
  - The speed and size of waves of different types of crime
  - Diffusion over a 2D (neighborhood) lattice with heterogeneous background prevalence.
    - **Is Crime Contagious?**

$$graffiti_{it} = \delta graffiti_{jt} + \beta X_{it} + \mu_t + \varepsilon_{it}$$

Keizer et al. *The Spreading of Disorder* – Science, Published 20 November 2008, 10.1126/science.1161405

COLLABORATORS: Joshua Plotkin, Caroline Buckee, Jon Wilkin, Jameson Toole

Summary of Some Open Questions…

- N = 1 HUNDRED
  - How to infer a relationships from many other temporal behavioral networks?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
  - How to identify the type of edge based on thousands of contextually labeled data points?

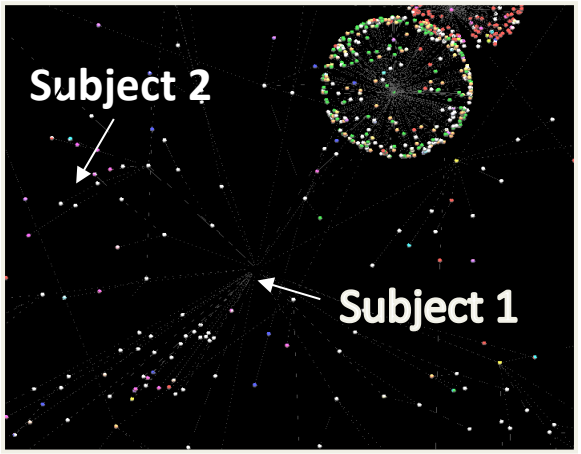
Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
  - How to leverage random sampling to learn about demographic groups?

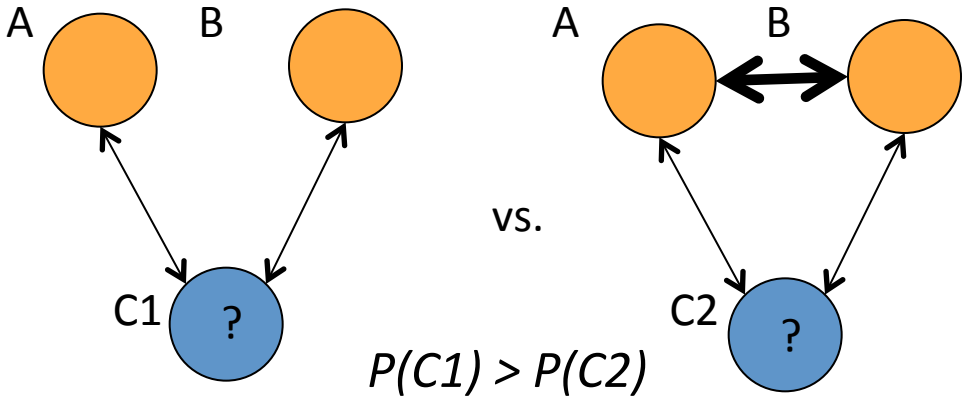
Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
  - How to disambiguate spread over a lattice with background prevalence?

Diffusion over the communication graph



- Does word-of-mouth diffusion follow similar dynamics in a rural Rwandan village as it does in London?
- How does region, socioeconomic status, phone type, impact adoption?



COLLABORATORS: Marcel Fafchamps (Oxford), Edo Airoldi (Harvard)

Location and Movements for Disease Dissemination

- How does season migrations impact computational epidemiological models of disease (malaria) eradication strategies?
- Is there a regional behavioral signature associated with a disease (cholera) outbreak?

COLLABORATORS: Neil Ferguson, Deirdre Hollingsworth, Azra Ghani, Jamie Griffin (Imperial)

Behavioral Reactions to Exogenous Events

- How does regional behavior change in relational to economic downturn (draught, crop prices, ...) or sudden prosperity (the IT sector in Kigali, …).
- How does regional behavior change in reaction to discrete shocks (earthquake, flooding, violence, …)

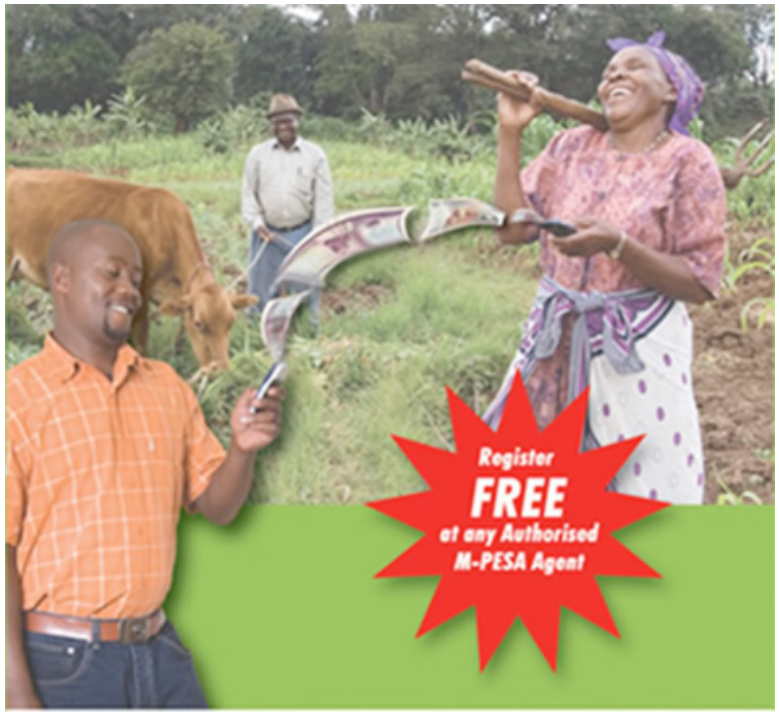
COLLABORATORS: Joshua Blumenstock (Berkeley)

N = 10 MILLION

Kenya: 85% of phones + location,
T = 3 years

Product Diffusion:

- Edges
  - Voice / SMS network
  - Flashing network
  - **Financial network**
- Nodes
  - Location
  - Product Adoption
  - Top-up History



COLLABORATORS: Betty Mwangi, Pauline Vaughan,(Safaricom), Marcel Fafchamps (Oxford)

N = 100 MILLION

UK: ~100% of landline and 80% of mobile phones, 12B edges, T = 1 month (08/05)

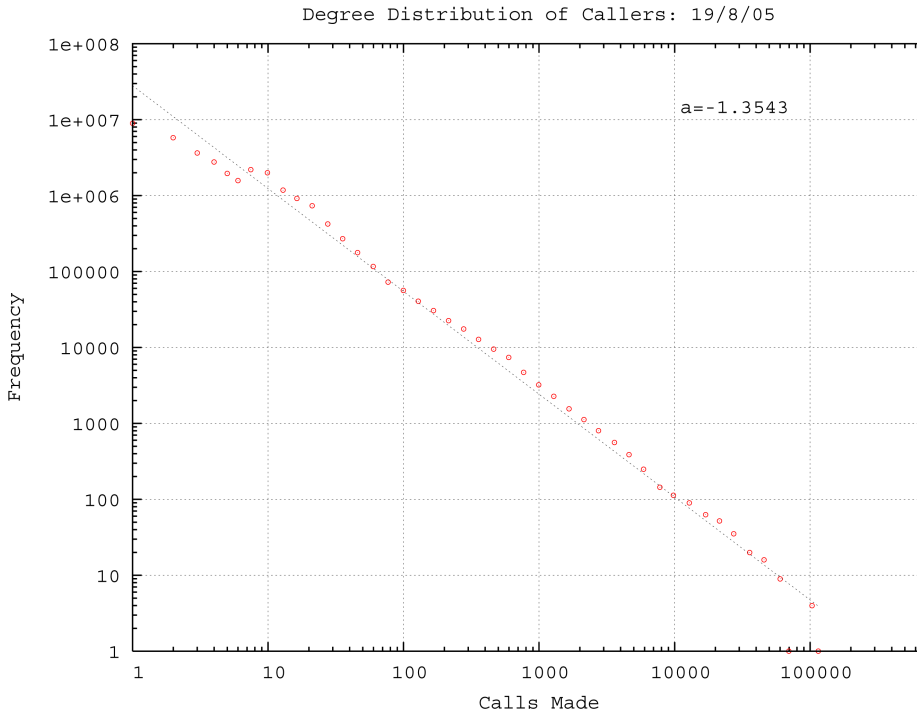
- Edge List
  - 250M nodes
  - 12B edges
- Nodes
  - Product Adoption
  - Area Code
- Edges
  - Time
  - Duration

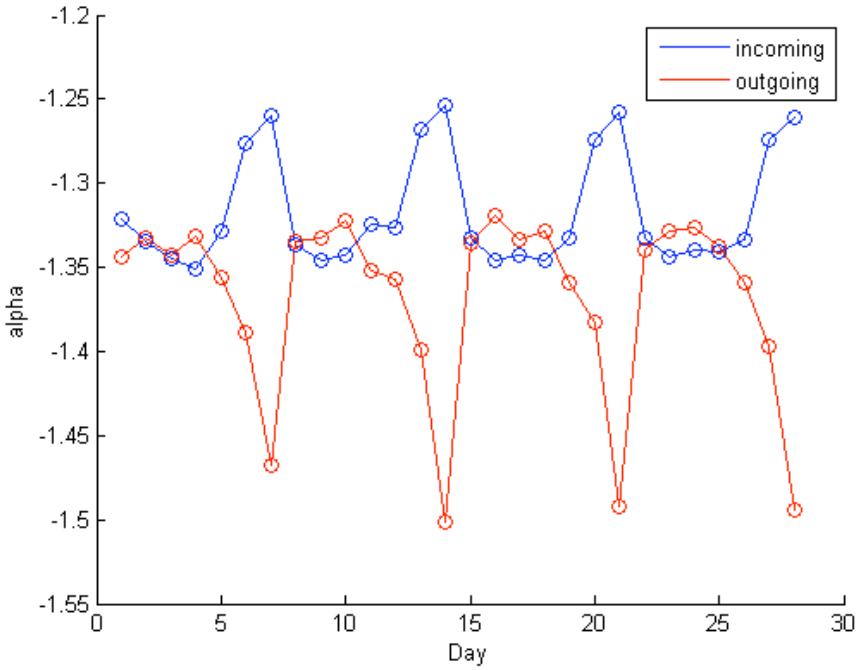
COLLABORATORS: Rob Claxton (BT)

N = 100 MILLION

UK: ~100% of landline and 80% of mobile phones, 12B edges, T = 1 month (08/05)

- What is driving degree distribution dynamics?

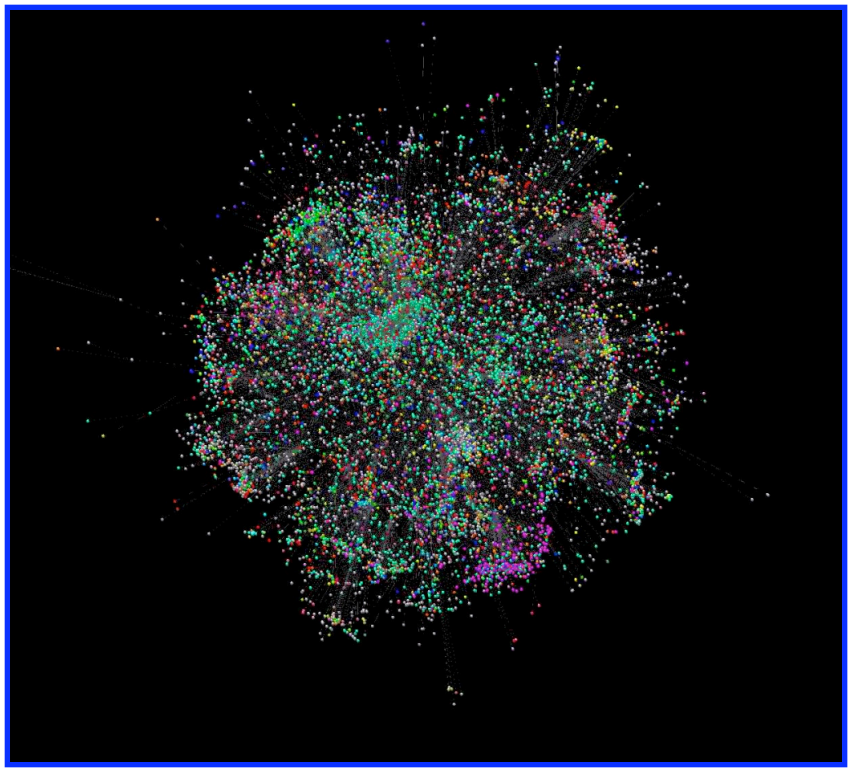




N = 100 MILLION

UK: ~100% of landline and 80% of mobile phones, 12B edges, T = 1 month (08/05)

- Graph Traversal on a 225M node giant component?
  - Binary Search:
    - *O(log2(N))*
    - *<50 ms lookup times / node*
  - *Parallel Binary Search*
    - *Raid 10 + 8 core = 8x speed-up*
    - *<10 ms/node*
    - *Failure rate with 64GB RAM ~.03%*

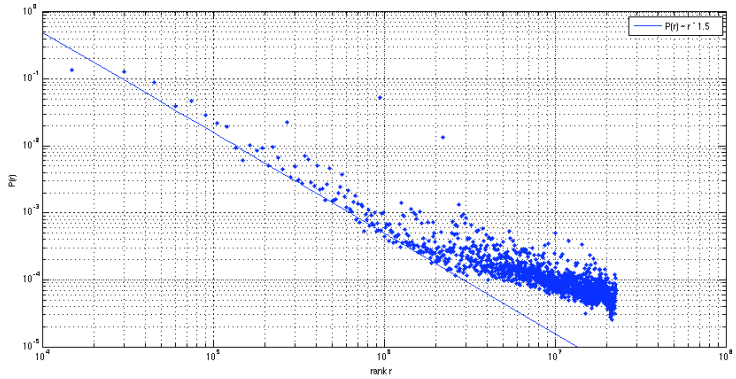


COLLABORATORS: Nate Metheny (SFI)

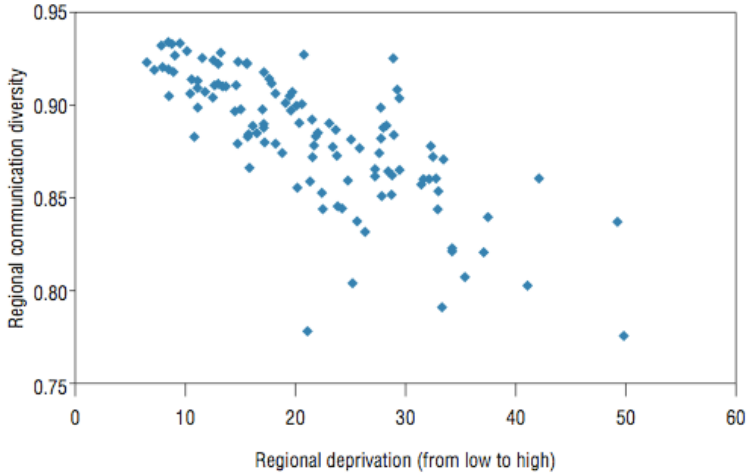
N = 100 MILLION

UK: ~100% of landline and 80% of mobile phones, 12B edges, T = 1 month (08/05)

- How does the probability of a tie scale with geographic distance?



- How is socio-economic status reflected in the call graph topology? Causation?



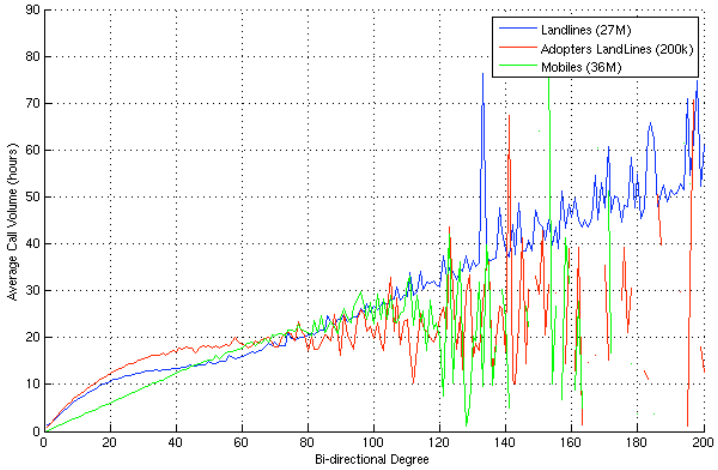
COLLABORATORS: Michael Macy (Cornell)

N. Eagle. 'Behavioral Inference Across Cultures: Using Telephones as a Cultural Lens', *IEEE Intelligent Systems*, Aug. 2008, Vol 23 (4), pp. 60-62.

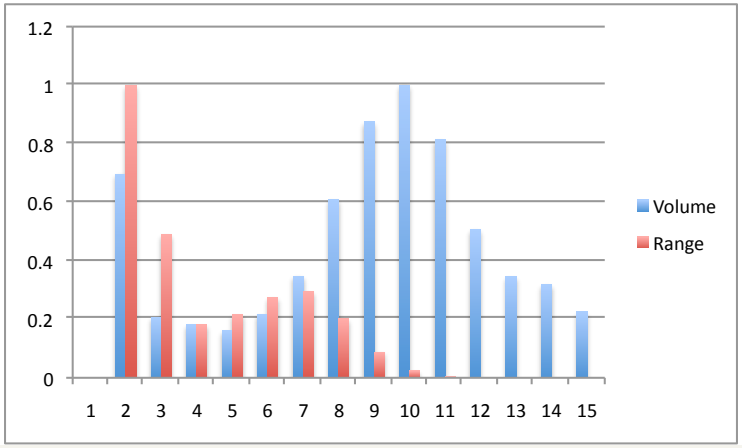
N = 100 MILLION

UK: ~100% of landline and 80% of mobile phones, 12B edges, T = 1 month (08/05)

- Categorizing Businesses vs. Residential Numbers



- Strength of Weak Ties



COLLABORATORS: Michael Macy (Cornell)

N = BILLIONS...

Bolivia, Dominican Republic, United States, Japan, Belgium, Thailand, Rwanda, United Kingdom, Kenya, Uganda, Saudi Arabia, Kuwait, India, Burkina Faso, Chad, Bahrain, Iraq, Jordan, Kuwait, Lebanon, Brazil, Spain, Saudi Arabia, DRC, Gabon, Ghana, Ireland, Madagascar, Malawi, Niger, Nigeria, Sierra Leone, Sudan, Tanzania, Uganda, Zambia

- The Social Network of Nations?
- Cultural Covariates
- Early warning of disease outbreaks / natural disasters?
- How to make life better?

COLLABORATORS: you?

Summary of Some Open Questions…

- N = 1 HUNDRED
  - How to infer a relationships from many other temporal behavioral networks?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
  - How to identify the type of edge based on thousands of contextually labeled data points?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
  - How to leverage random sampling to learn about demographic groups?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
  - How to disambiguate spread over a lattice with background prevalence?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
  - How is recent urbanization affecting people's support networks?
  - How can we better understand disease dynamics with actual mobility patterns?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
- N = 10 MILLION
  - How do resources flow through social networks?

Summary of Some Open Questions…

- N = 1 HUNDRED
- N = 1 THOUSAND
- N = 10 THOUSAND
- N = 100 THOUSAND
- N = 1 MILLION
- N = 10 MILLION
- N = 100 MILLION
  - What is driving the behavior of the aggregate?
  - Strength of weak ties?
  - Graph Traversal Using Parallel Binary Search on Sorted Edge Lists?