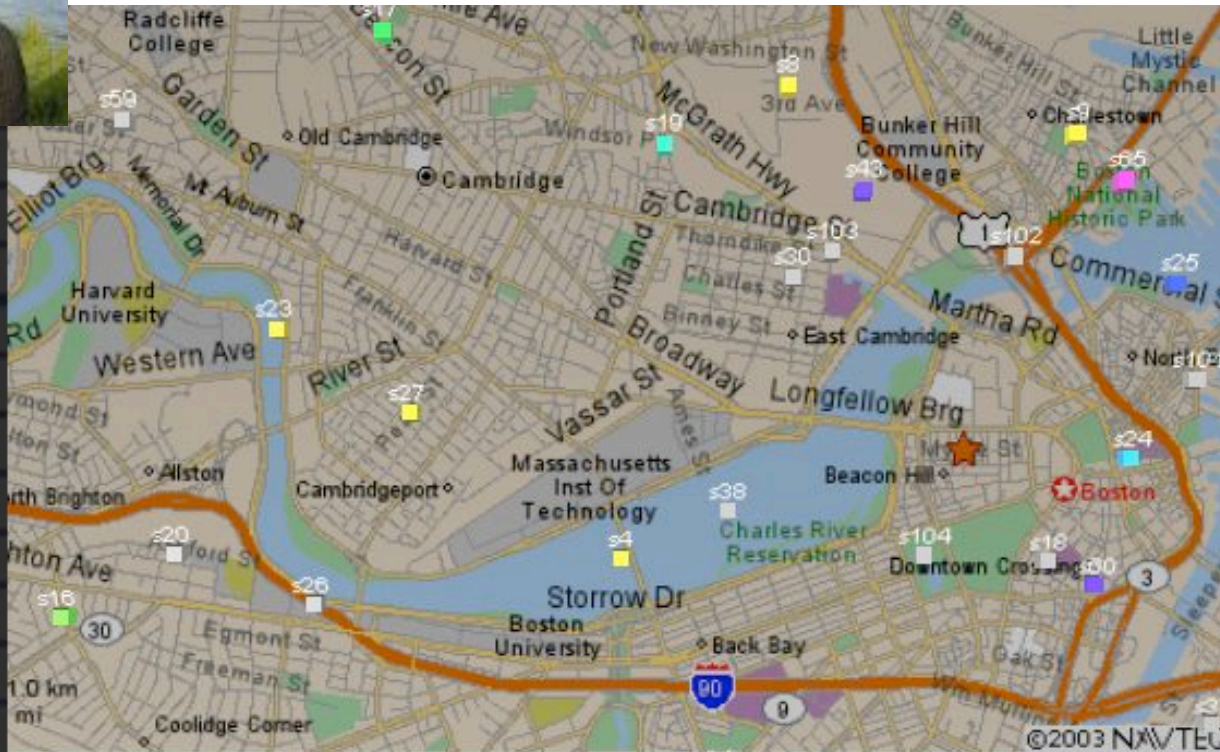


Open Questions from Large-Scale Behavioral Data Analysis

Scaling from hundreds, to thousands,
to millions, to (potentially) billions.

1 HUNDRED



Advisor Group

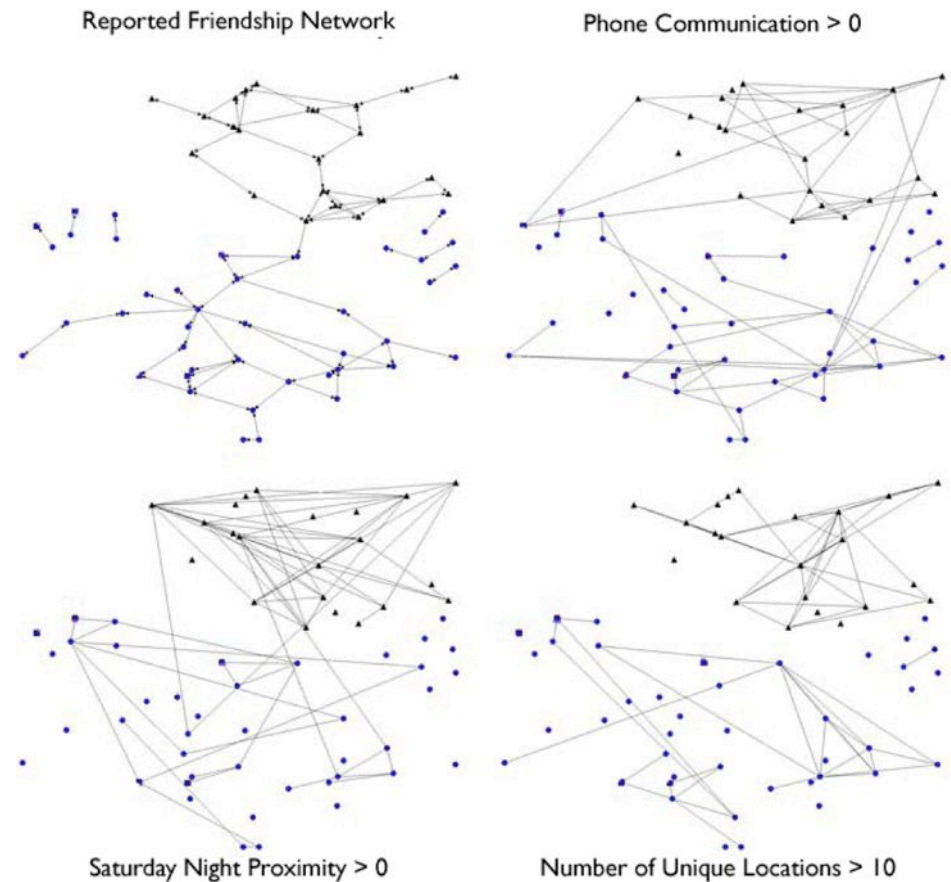
mitch sandy pattie chris neenys chriss chriso bill henry joej barry hugh marvin



1 HUNDRED

SOME OPEN QUESTIONS

- Dynamics of Proximity Networks?
- Generative Models?
 - context, temporal, spatial
- Inferring Friendship?



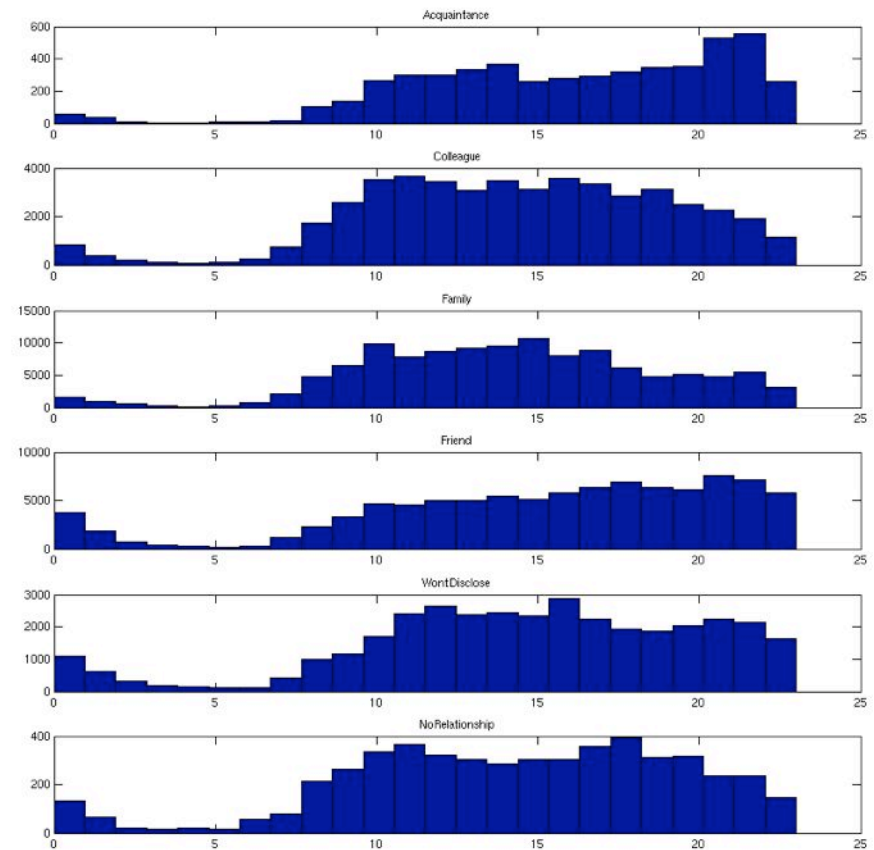
COLLABORATORS: David Lazer, Sandy Pentland, Aaron Clauset



1 THOUSAND

SOME MORE OPEN QUESTIONS

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



COLLABORATORS: Hannu Verkasalo, Cosma Shalizi, Alessandro Rinaldo, Raja Hafiz



10 THOUSAND

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



COURTESY OF SENSENETWORKS: Greg Skibiski, Tony Jebara

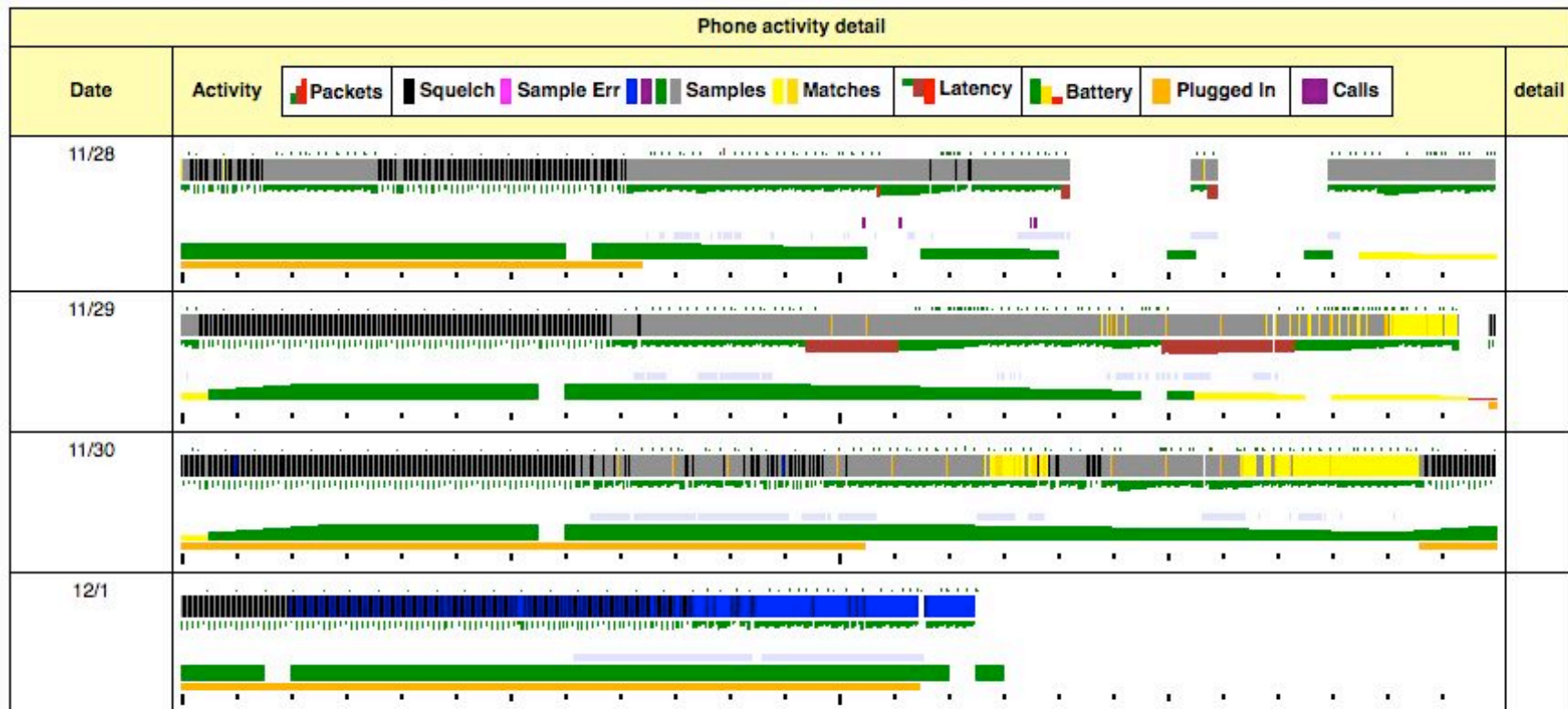


10 THOUSAND



IMMI Media Monitoring System Activity Report -/-

Panelist: 12254
Market: 7 Los Angeles (PST8PDT)



Times in PST8PDT
Offset 8 hours

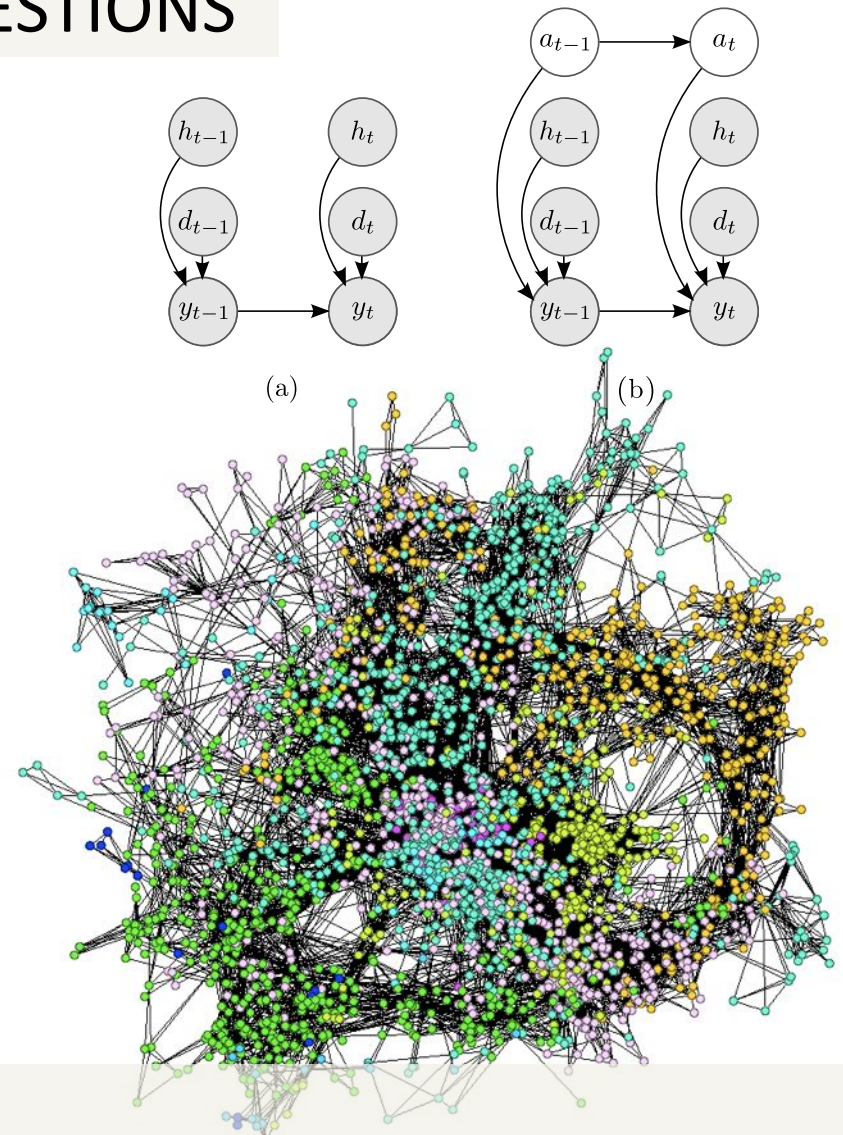


10 THOUSAND

SOME OPEN QUESTIONS

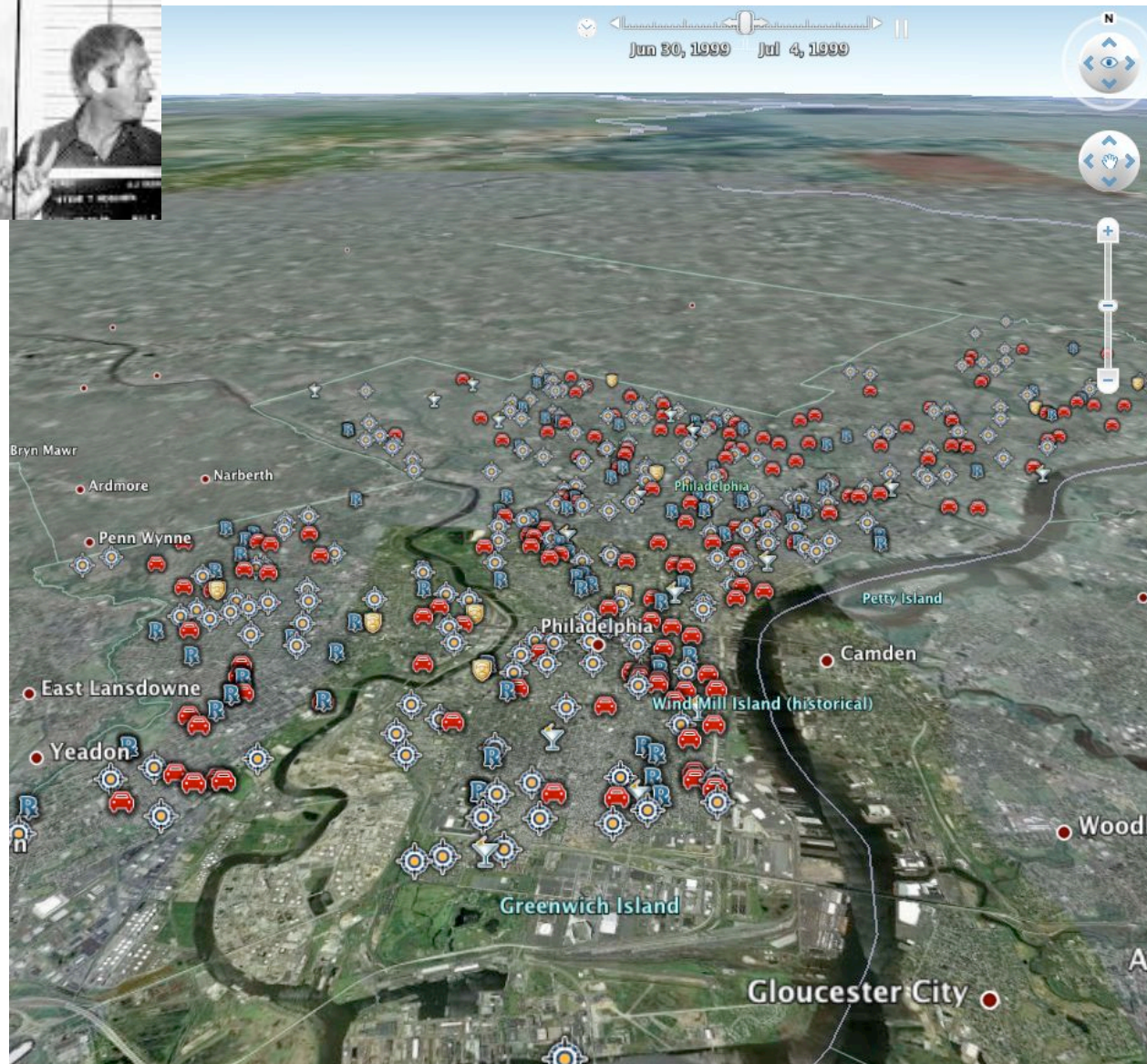
- How to scale DBNs to much higher dimensional state-space?
- Detect Outlier Events?
- Demographic Segmentation?

demographic (N)	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

100 THOUSAND

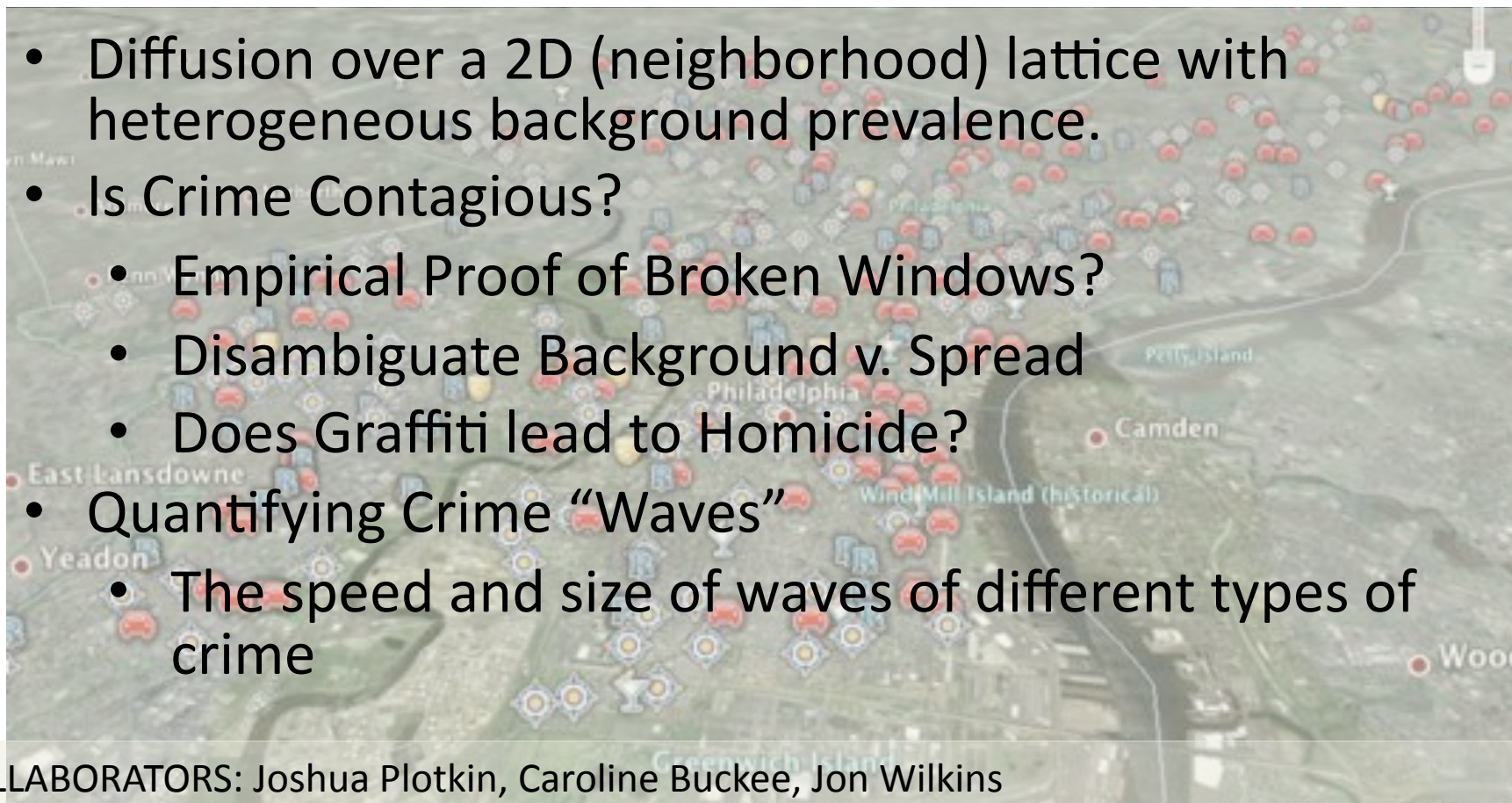




100 THOUSAND

SOME OPEN QUESTIONS

- Diffusion over a 2D (neighborhood) lattice with heterogeneous background prevalence.
- Is Crime Contagious?
 - Empirical Proof of Broken Windows?
 - Disambiguate Background v. Spread
 - Does Graffiti lead to Homicide?
- Quantifying Crime “Waves”
 - The speed and size of waves of different types of crime



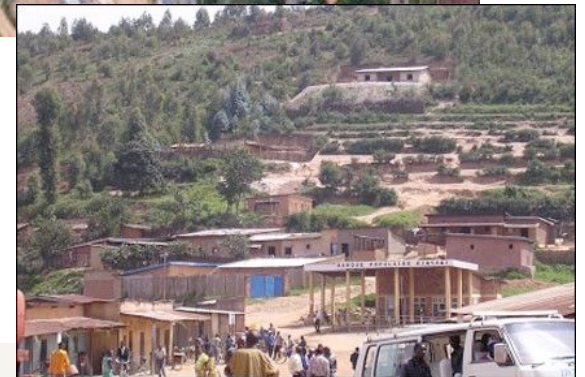
COLLABORATORS: Joshua Plotkin, Caroline Buckee, Jon Wilkins



1 MILLION

SOME OPEN QUESTIONS

- Co-evolution of Social Networks and Cities
 - Dynamics of tie formation (& slums)?
 - Which ties (neighborhoods) persist?
 - How does mobility affect topological structure?
 - Disease ramifications?
 - Census Data...



COLLABORATORS: Marta Gonzales, Luis Bettencourt



10 MILLION

“Next Generation” Behavioral Data

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

New Set of Tools Needed!



COLLABORATORS: Betty Mwangi, Pauline Vaughan, Marcel Fafchamps



100 MILLION



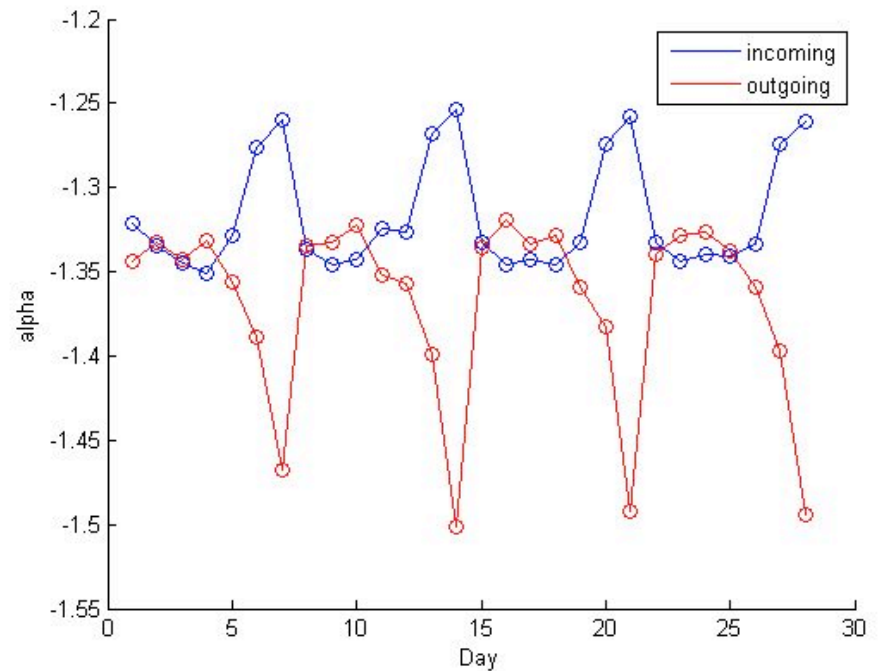
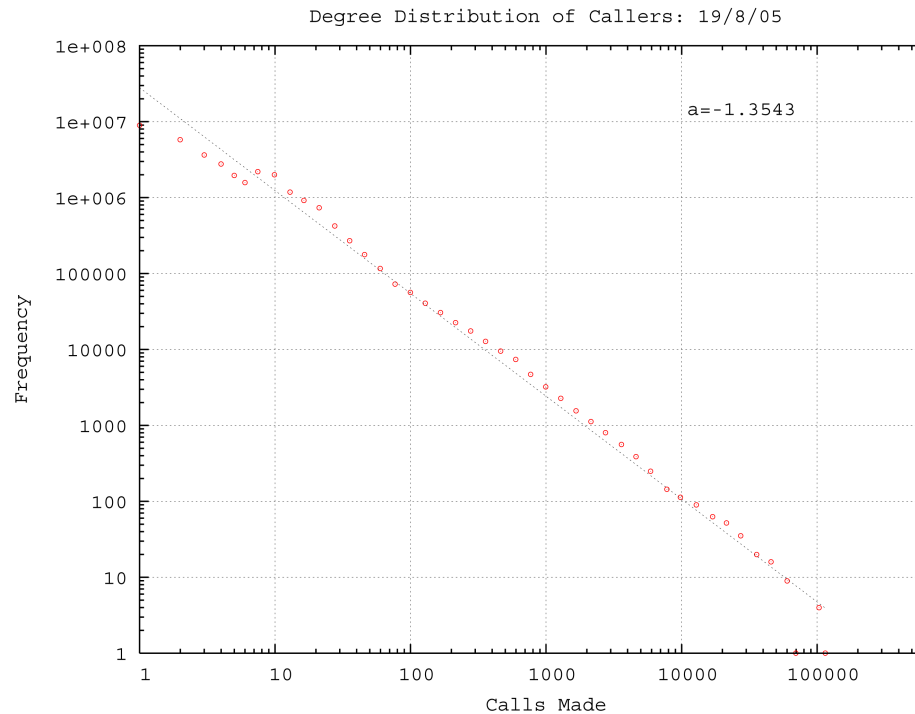
COLLABORATORS: Rob Claxton, Jon Reades



100 MILLION

SOME OPEN QUESTIONS

- What is driving degree distribution dynamics?



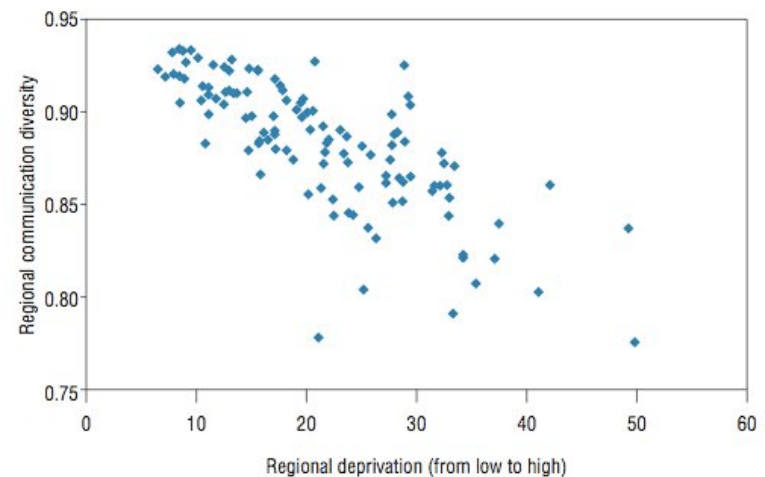
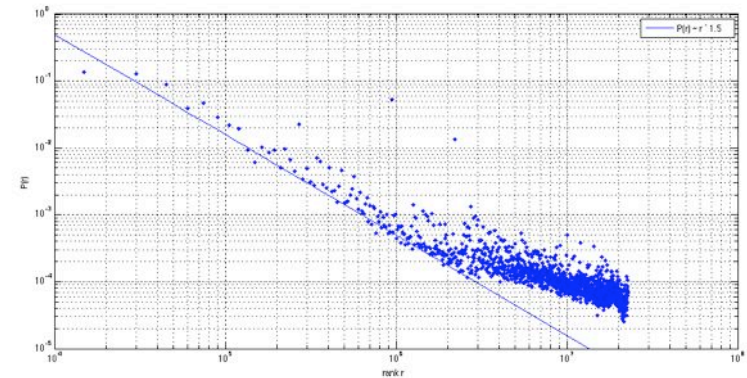
COLLABORATORS: Michael Macy



100 MILLION

SOME OPEN QUESTIONS

- How does the probability of a tie scale with geographic distance?
- How is socio-economic status reflected in the call graph topology?
Causation?

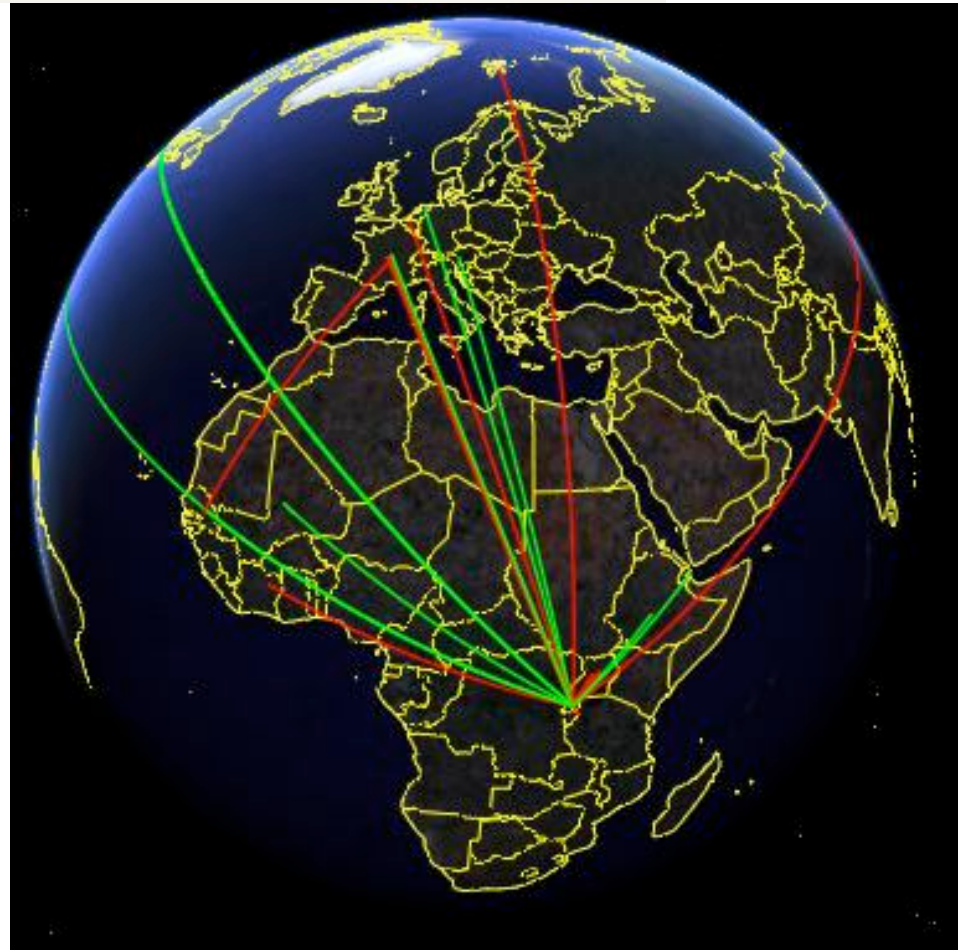




BILLIONS...

LOTS AND LOTS OF OPEN QUESTIONS

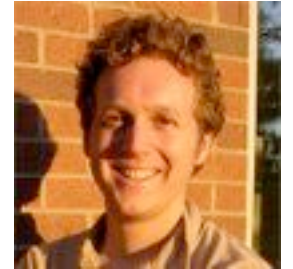
- The Social Network of Nations?
- Cultural Covariates
- Early warning of disease outbreaks / natural disasters?
- How to make life better?



COLLABORATORS: you?

Summary of Some Open Questions...

- 1 HUNDRED
 - How to infer an edge from many other behavioral networks?



Summary of Some Open Questions...

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



Summary of Some Open Questions...

- 1 HUNDRED
- 1 THOUSAND
- 10 THOUSAND
 - How to leverage random sampling to learn about demographic groups?
 - Which ones are likely to end up in the hospital?



Summary of Some Open Questions...

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



Summary of Some Open Questions...

- 1 HUNDRED
- 1 THOUSAND
- 10 THOUSAND
- 100 THOUSAND
- 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...

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



Summary of Some Open Questions...

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



Summary of Some Open Questions...

- 1 HUNDRED
- 1 THOUSAND
- 10 THOUSAND
- 100 THOUSAND
- 1 MILLION
- 10 MILLION
- 100 MILLION
- 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?

thanks.