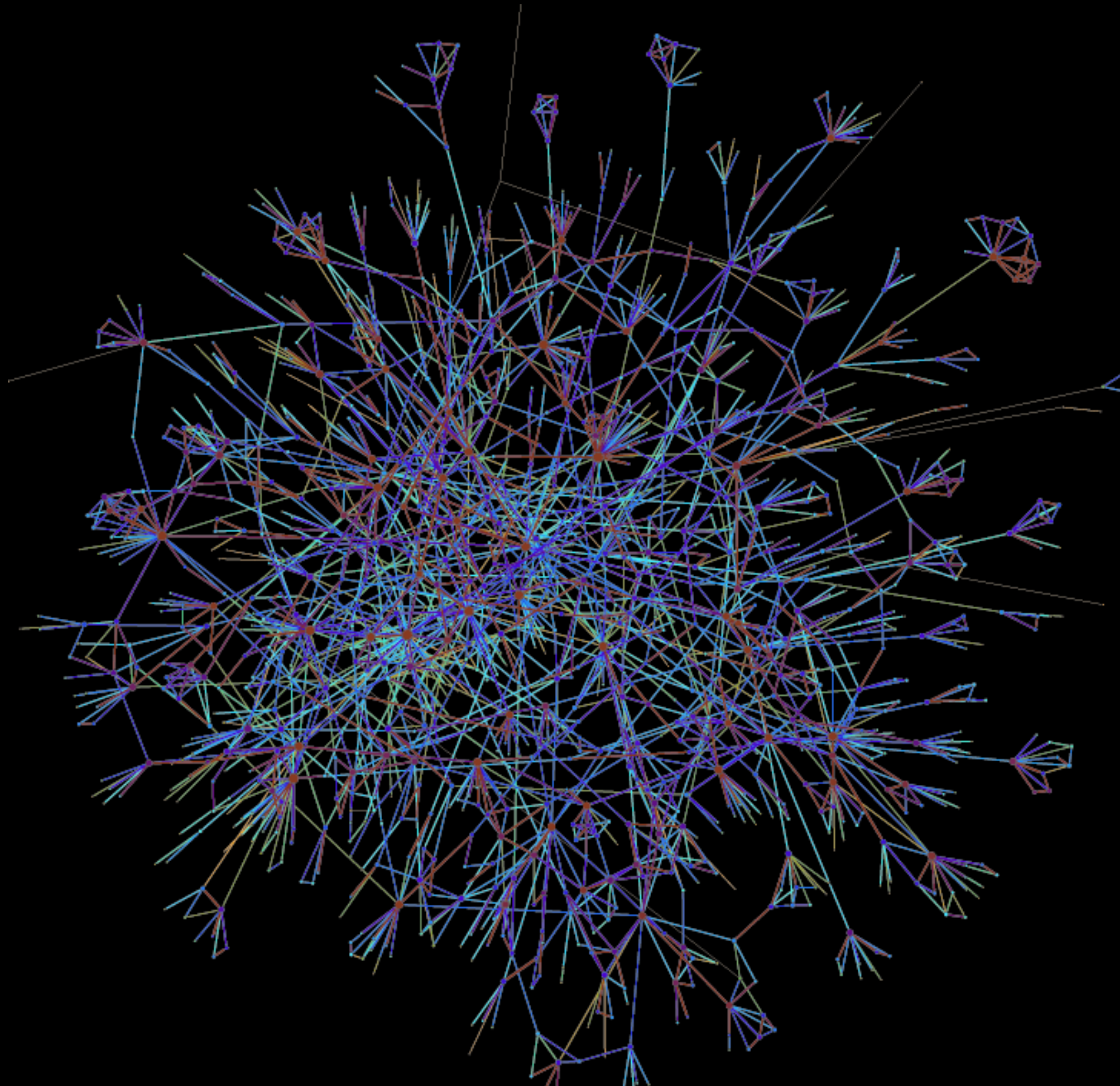
Life in the network

The coming age of computational social science



Computational social science

Computational social science

- The capturing and analysis of human activity represented in digital form
  - Increased computational capacity to manipulate data
  - Incidental, vast archives of human activity (e.g., Internet, e-mail)
  - Instrumentation of human behavior (e.g., cookies, GPS devices)
  - Creation of virtual worlds to experiment with
- What are the implications for our understanding of collective human behavior?
- What are the obstacles to the emergence of a “computational social science”?

What can data like these tell us?

- How do “things” spread through a network?
  - Ideas?
  - Avian flu?
- How do people/organizations work together?
  - Collaboration and coordination?
  - Who is in key positions in the network?
- Form an empirical basis for various types of policy recommendations
- Possibly even real-time feedback for effective interventions

Computational social science

- Orders of magnitude increase in data being collected about human behavior over last decade
- Constant increase in computational power
- Shift in social science research over the next generation
  - Thinking relationally: what is flowing among people? How are people working together?

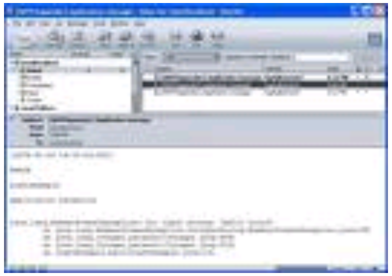
Program on Networked Governance

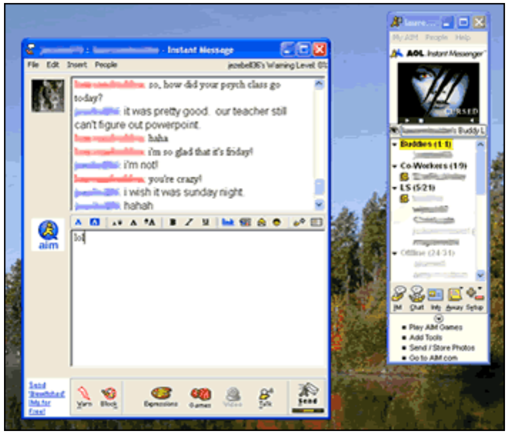
Existing approaches to studying networks

- Exponential growth in social network analysis in last decade, especially in the study of organizations, but generally across the academy
- In social sciences, generally rely on self reports
- **Static**– generally based on snapshots
- **Shaky reliability**– what is being measured by self reports?
- **Small scale**– mostly systems in the hundreds or less
- → Inferential challenges in existing research
- → Many important phenomena are neglected

Program on Networked Governance

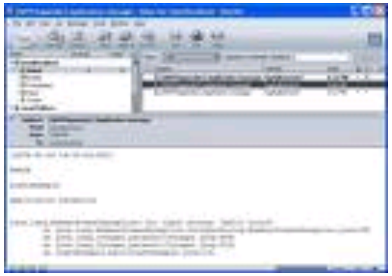
Life in the network

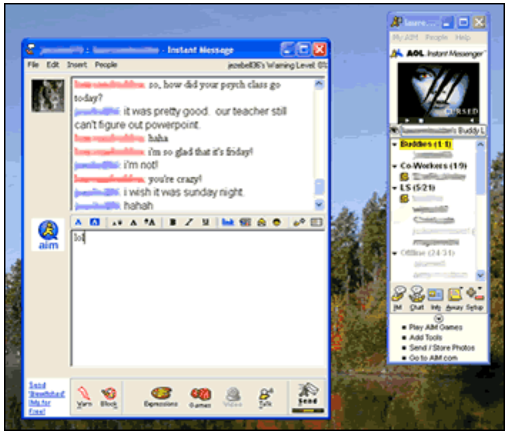




Life in the network

- E-mail
- Instant messaging
- Text messaging
- Telephone logs
- Link structure among websites (google algorithm)
- References (e.g., social science index)
- ***What can data like these tell us?***

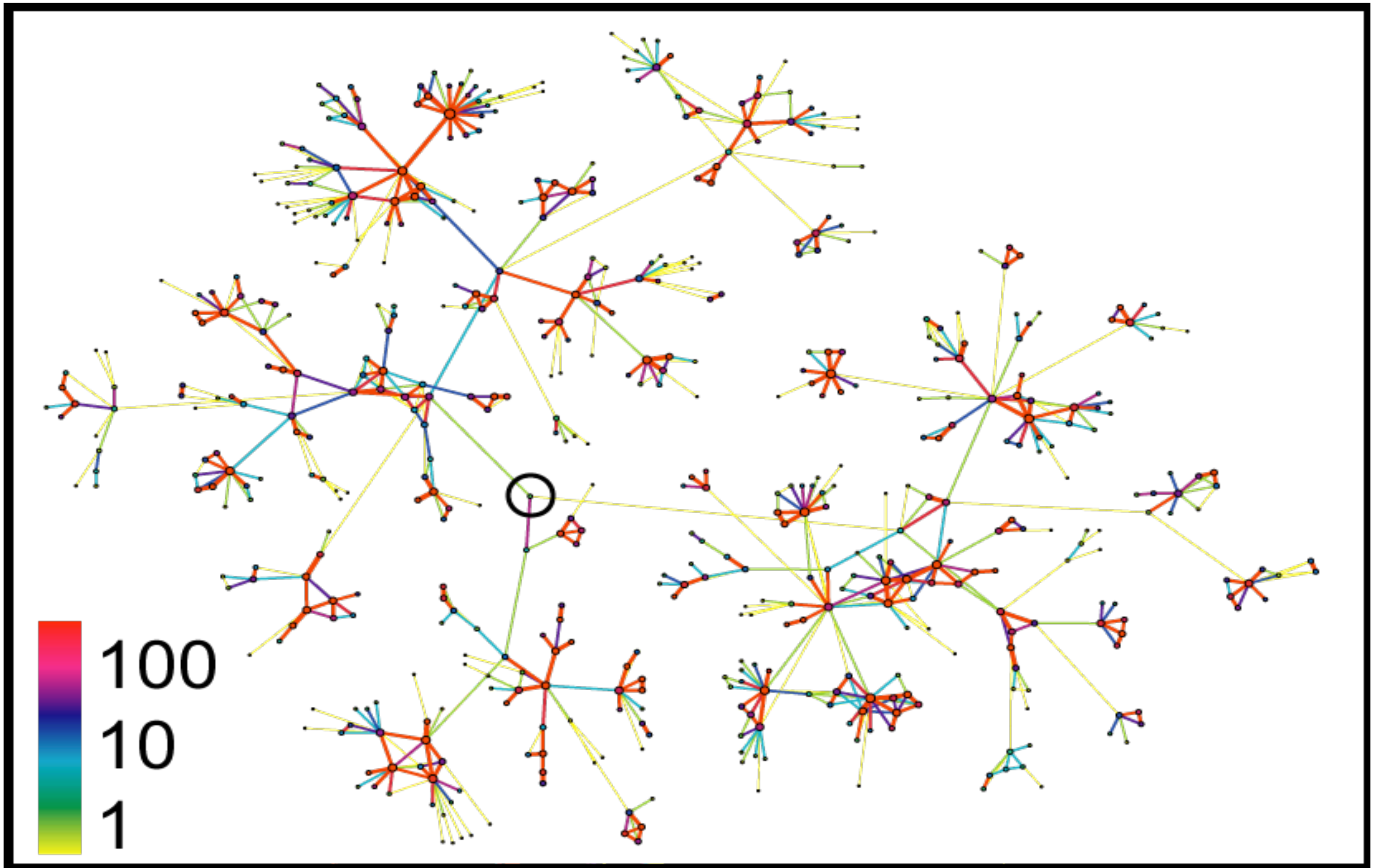




Study 1: Call log analysis

- "Structure and tie strengths in mobile communication networks" (*PNAS, with J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó,K. Kaskil, J. Kertész, A.-L. Barabási*)
- Examination of call log data from mobile phone company in moderate sized European nation– a total of approximately **7,000,000** users, 49 trillion dyads
- What does network structure look like?

Call log network data



Results…

- Hub-spoke structure (scale free)
- Small world (on average, 13 degrees of separation)
- But poorly structured for dissemination: Strong ties tend to be clustered, and weak ties bind clusters together (consistent with Granovetter)
- But simulations suggest that weak(est) ties are *not* effective at spreading (inconsistent with Granovetter)
- Potentially powerful tool for studying evolving social structures of communities
- Possible use of data for a variety of policy purposes, from criminal investigations to “early warning” system for avian flu

- *But: what does a phone call between two people mean??*

Study 2: Instrumentation of human behavior

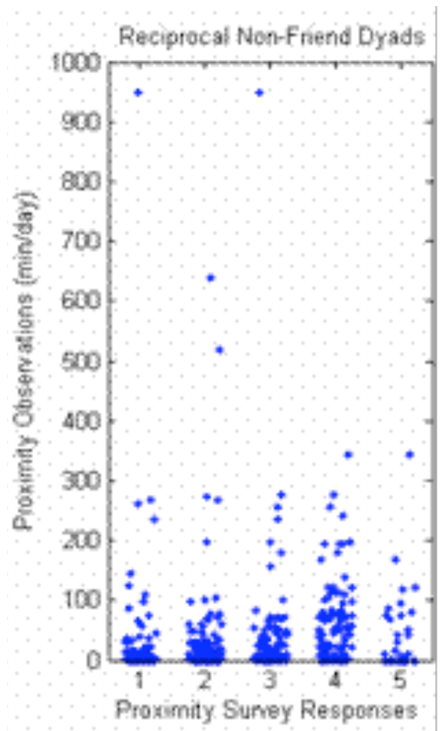
- Paper: "Revealing Social Relationships using Contextualized Proximity and Communication Data" (with Nathan Eagle and Sandy Pentland)
- Collaboration with Media Lab
- Program mobile phones of ~100 students for 9 months:
  - Call log data
  - Physical proximity (using Bluetooth)
  - Location (using cell tower triangulation)
- Also collected self report data on friendship, satisfaction
- What is the information in these data?
- Compare observations to self reports

Self reported vs observed proximity

- Substantial recency effects: recent interactions weighted more heavily
- Reciprocal non-friends: 99.5% accurate at reporting 0’s
- Reciprocal friends: 35% accurate at reporting 0’s
- Friends more accurate at non-0’s

Self reported vs observed proximity

- Substantial recency effects: recent interactions weighted more heavily
- Reciprocal non-friends: 99.5% accurate at reporting 0’s
- Reciprocal friends: 35% accurate at reporting 0’s
- Friends more accurate at non-0’s



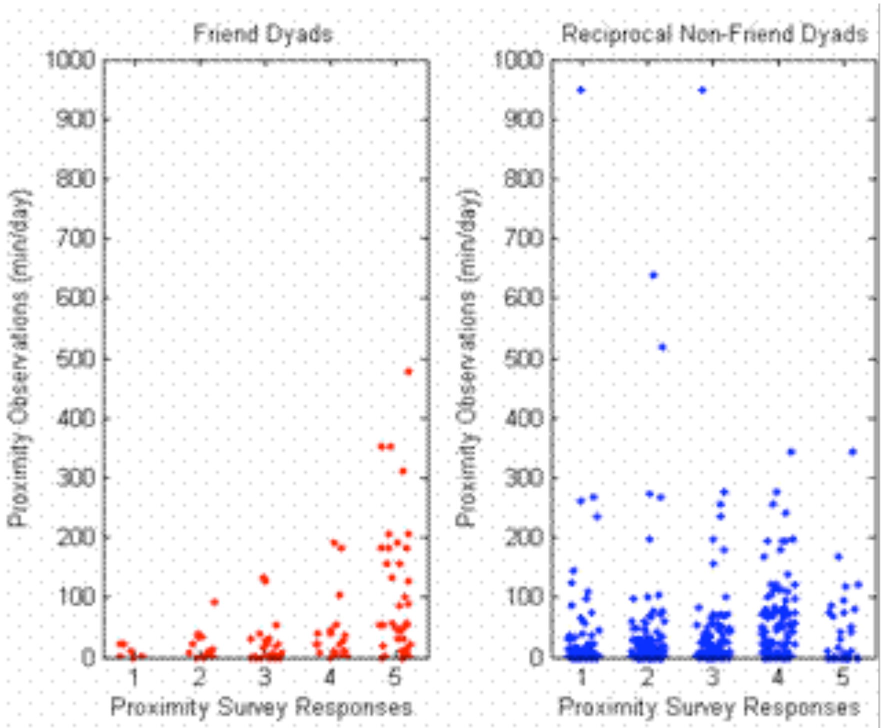
Survey Key

Response	-- Reported Avg. Proximity
0	< 5 minutes
1	5-10 minutes
2	10-30 minutes
3	30 minutes - 2 hours
4	2-4 hours
5	> 4 hours

Program on Networked Governance

Self reported vs observed proximity

- Substantial recency effects: recent interactions weighted more heavily
- Reciprocal non-friends: 99.5% accurate at reporting 0's
- Reciprocal friends: 35% accurate at reporting 0's
- Friends more accurate at non-0's



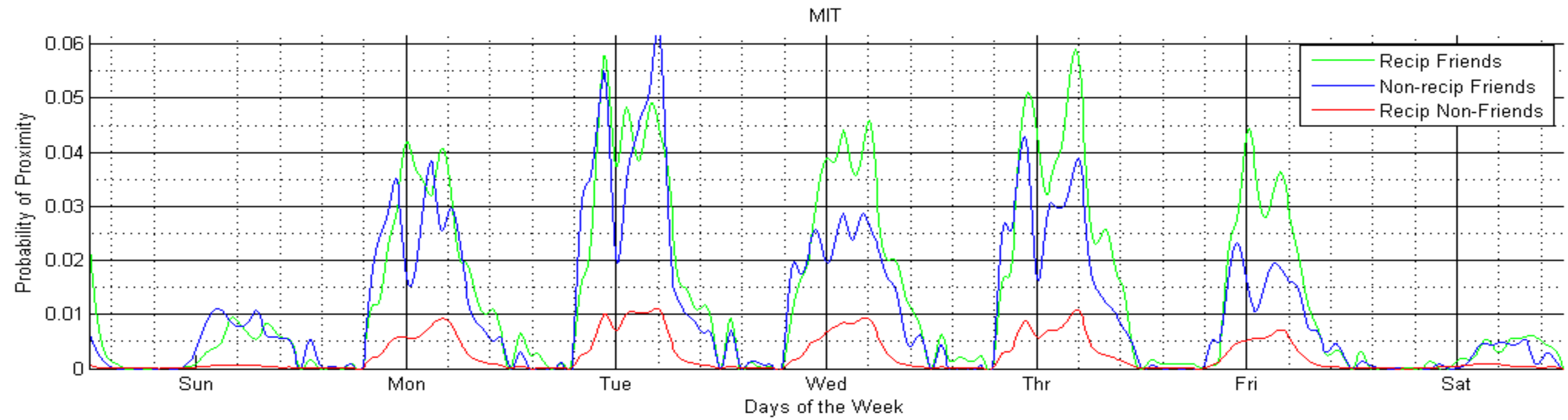
Survey Key

Response	-- Reported Avg. Proximity
0	< 5 minutes
1	5-10 minutes
2	10-30 minutes
3	30 minutes - 2 hours
4	2-4 hours
5	> 4 hours

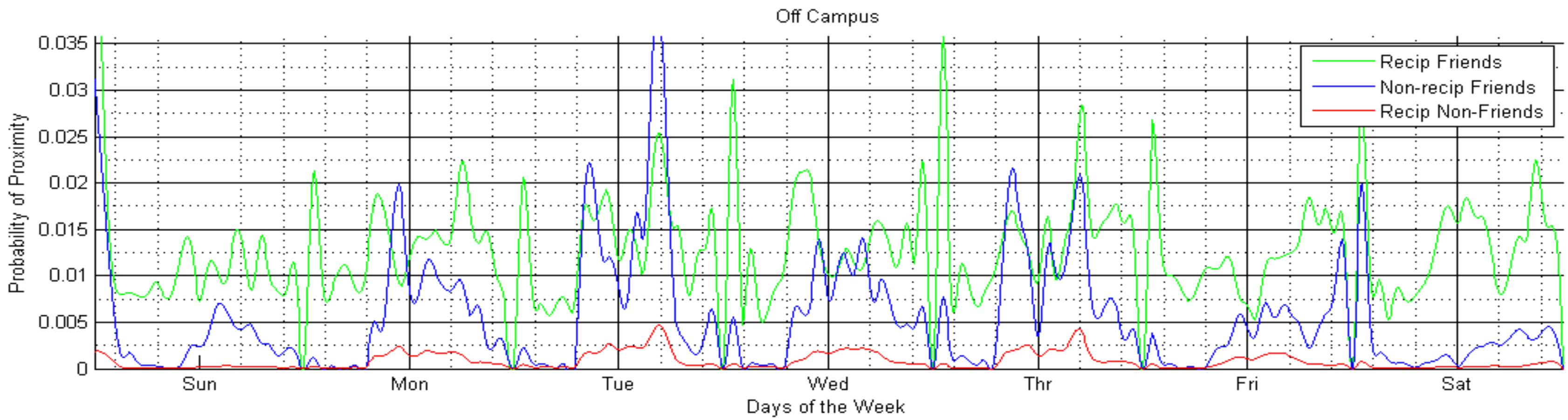
Program on Networked Governance

Is friendship observable?

- Friendship is important at individual and collective levels due to the resources that flow among friends
- "Purely" cognitive relationship: in principle, you could be friends with someone with whom you do not interact.
- But generally we all make inferences about who is friends with whom based on our observations
- Can the types of information that inform our inferences be captured via our mobile phones?
  - Certainly, one anticipates that (for ex) friends will tend to be proximate to each other
- *If high accuracy is possible, then possible to look at evolution of friendship structure in larger populations over time (as well as other cognitive relationships, such as advice)*



MIT
0.06
0.05
0.04
0.03
0.02
0.01
0
Probability of Proximity
Sun
Mon
Tue
Wed
Thr
Fri
Sat
Days of the Week
Recip Friends
Non-recip Friends
Recip Non-Friends



Off Campus
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Probability of Proximity
Sun
Mon
Tue
Wed
Thr
Fri
Sat
Days of the Week
Recip Friends
Non-recip Friends
Recip Non-Friends

Self reported versus observed friendships

- We were able to categorize correctly ***~95%*** of reciprocated friendships and reciprocated non-friendships with a single parameter
- Unreciprocated "friendships" came from high scores in-role communication, perhaps capturing cultural ambiguity
- Created continuous construct from dichotomous self report– perhaps a more valid measure of friendship?
- Second layer of validation: predicting satisfaction based on (a) actual friendships and (b) inferred friendship. Second model does slightly better.
- ***Results suggest potential for inferring friendship on much larger scale.***

The future…

- Bridging narrow and deep versus broad and shallow data collections…
  - Scale up “deep” data collection
  - Build capacity to infer deeper things about “shallow” data
- Development of designs that match unique characteristics of data: quantity of data is no substitute for quality of design
- Examine substantive phenomena within network, evolution of friendship structures, social capital and demography, epidemiology, etc

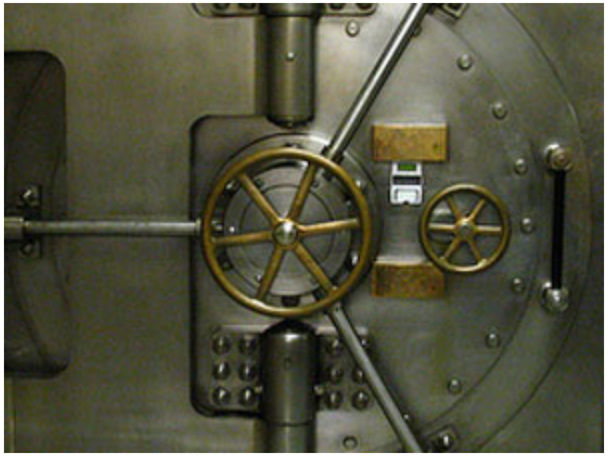
<u>HUGE</u> privacy issues

- Many (not all) of these data involve privacy concerns
- Examples: movement, e-mail data, instant messaging, etc.
- Current model of “let 1000 flowers bloom” is good for innovation, bad for potential privacy breaches
- IRBs are generally not savvy to all of the ways data can be de-anonymized
- Ex of recent pulling down of NIH data

Program on Networked Governance

<u>HUGE</u> data access issues

- Two possible dystopias
  - CSS remains the domain largely of corporations and government agencies
  - Dead sea scrolls model, where researchers gain access to data, but don't share



And major institutional challenges for the academy…

- Overcoming silo's of academia, particularly wide between the sciences and social sciences

Program on Networked Governance

Issues to think about:

- **What network data are meaningful?** There are potentially serious heterogeneity with behavioral data.
- **When are strong statistical regularities interesting?** [often, they are not so interesting]
- **When are large quantities of data valuable?** (as compared to small, high quality, samples)
  - Russian military proverb: *Quantity has its own quality.*

Program on Networked Governance


The Whirlpool Galaxy — M51
HUBBLESITE.org