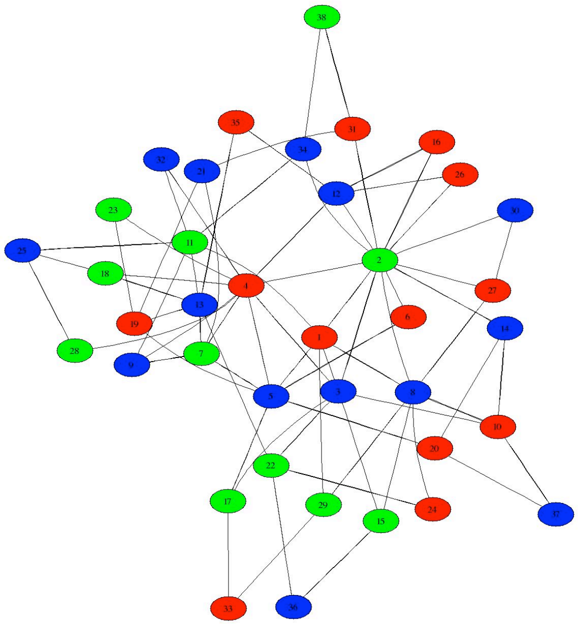
Structural Properties of Networks: Introduction

Networked Life
NETS 112
Fall 2019
Prof. Michael Kearns

Networks: Basic Definitions

- A network (or graph) is:
  - a collection of individuals or entities, each called a vertex or node
  - a list of pairs of vertices that are neighbors, representing edges or links
- Examples:
  - vertices are mathematicians, edges represent coauthorship relationships
  - vertices are Facebook users, edges represent Facebook friendships
  - vertices are news articles, edges represent word overlap
- Networks can represent any binary relationship over individuals
- Often helpful to visualize networks with a diagram
- But to us, the network is the list of edges, not the visualization
  - same network has many different visualizations

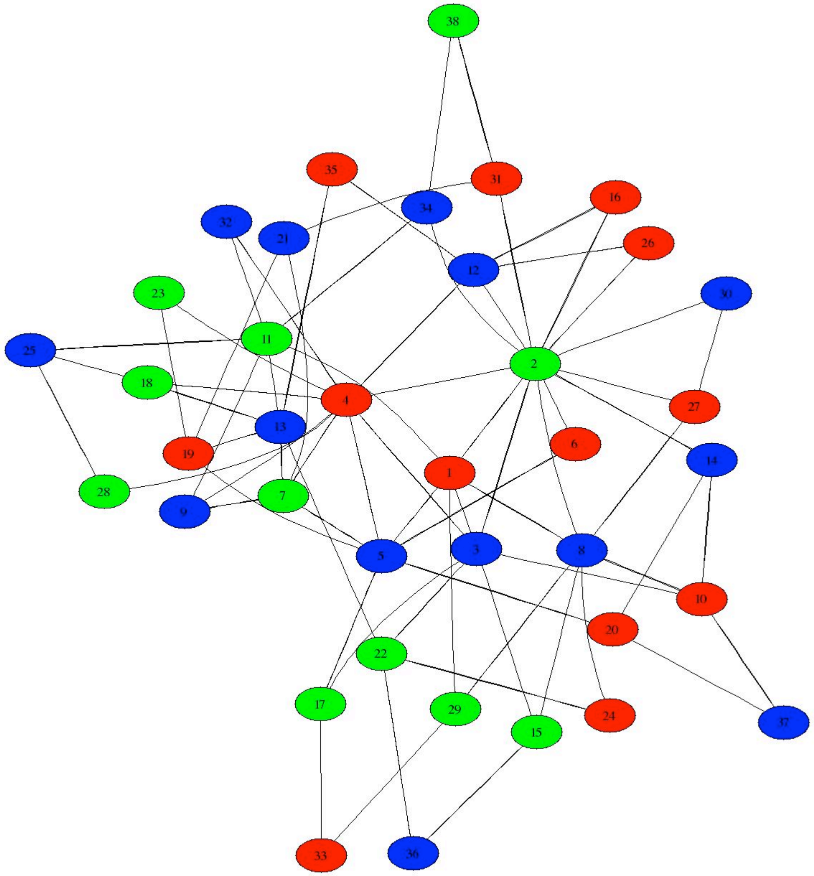


Networks: Basic Definitions

- We will use N to denote the number of vertices in a network
- Number of possible edges:

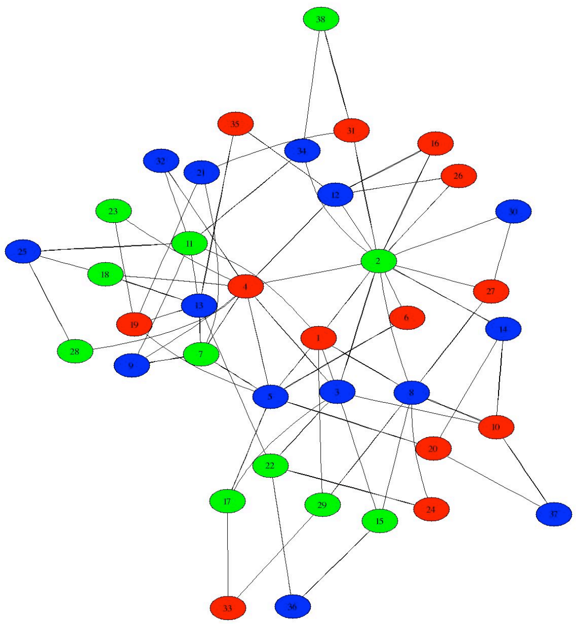
$$N(N-1)/2 \approx N^2/2$$

- The degree of a vertex is its number of neighbors



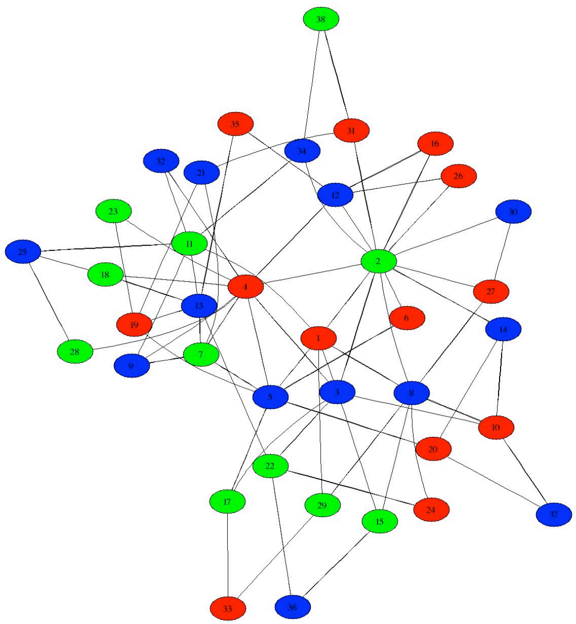
Networks: Basic Definitions

- The distance between two vertices is the length of the shortest path connecting them
- This assumes the network has only a single component or “piece”
- If two vertices are in different components, their distance is undefined or infinite
- The diameter of a network is the average distance between pairs
- It measures how near or far typical individuals are from each other



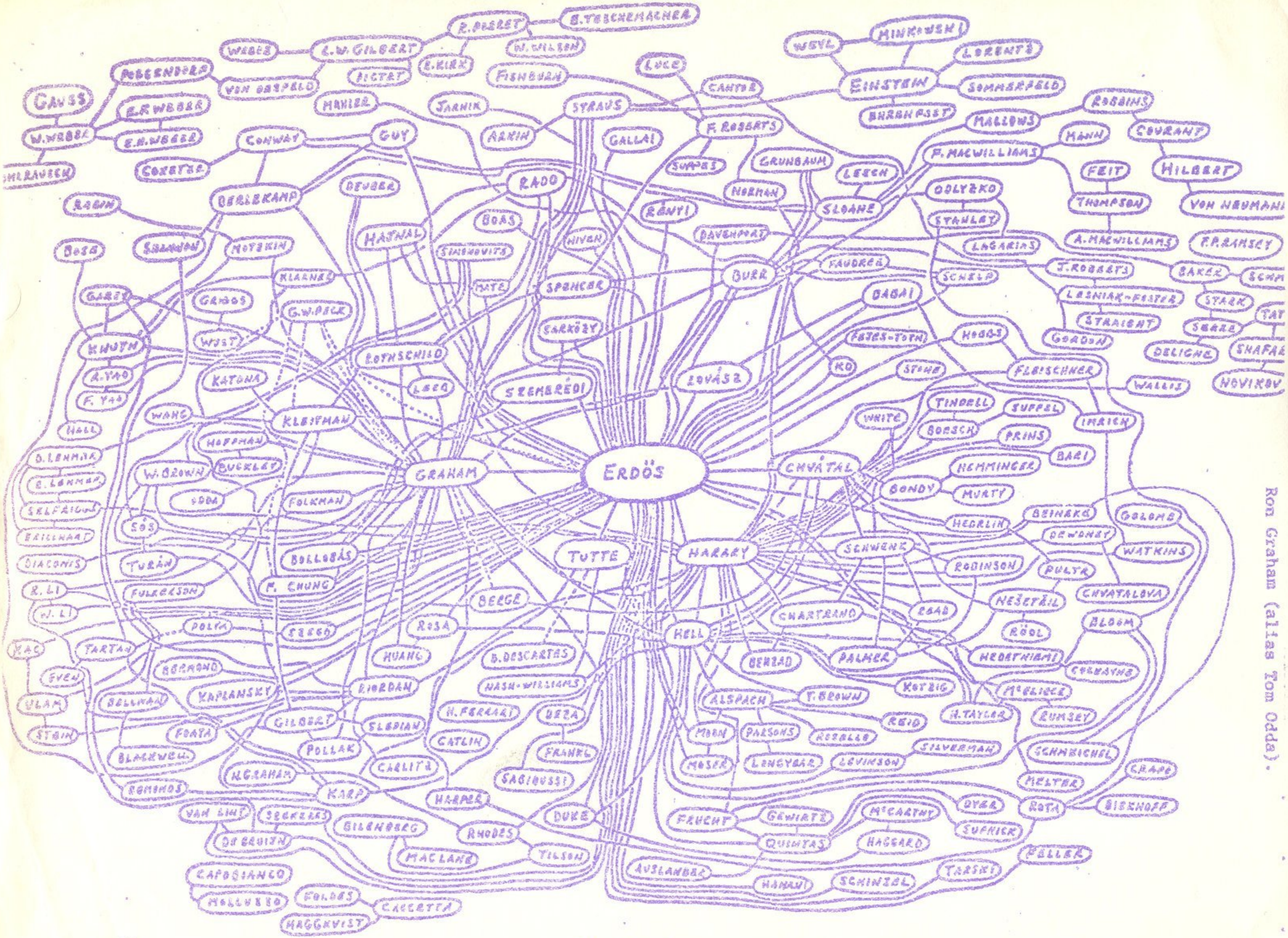
Networks: Basic Definitions

- So far, we have been discussing undirected networks
- Connection relationship is symmetric:
  - if vertex u is connected to vertex v, then v is also connected to u
  - Facebook friendship is symmetric/reciprocal
- Sometimes we'll want to discuss directed networks
  - I can follow you on Twitter without you following me
  - web page A may link to page B, but not vice-versa
- In such cases, directionality matters and edges are annotated by arrows



Illustrating the Concepts

- Example: scientific collaboration
  - vertices: math and computer science researchers
  - links: between coauthors on a published paper
  - *Erdos numbers* : distance to Paul Erdos
  - Erdos was definitely a *hub* or *connector;* had 507 coauthors
  - MK's Erdos number is 3, via Kearns → Mansour → Alon → Erdos
  - how do we *navigate* in such networks?
  - how does network distance relate to the real world?



Ron Graham (alias Tom Odda).

Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Erdös number 0	---	*1 person*
Erdös number 1	---	*504 people*
Erdös number 2	---	*6593 people*
Erdös number 3	---	*33605 people*
Erdös number 4	---	*83642 people*
Erdös number 5	---	*87760 people*
Erdös number 6	---	*40014 people*
Erdös number 7	---	*11591 people*
Erdös number 8	---	*3146 people*
Erdös number 9	---	*819 people*
Erdös number 10	---	*244 people*
Erdös number 11	---	*68 people*
Erdös number 12	---	*23 people*
Erdös number 13	---	*5 people*

The median Erdös number is 5; the mean is 4.65, and the standard deviation is 1.21.

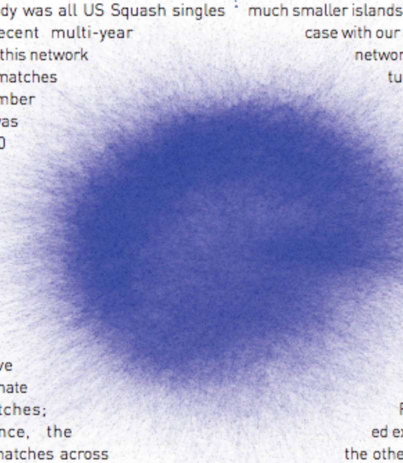
THE SMALL-WORLD NETWORK OF SQUASH

BY MICHAEL KEARNS AND RYAN RAYFIELD

Not all social networks are built in front of glowing monitors with a Mountain Dew and a bag of Cheetos at hand. There are some social networks in which participation is outright good for your health—like squash. Using tools from the emerging field of network science we will investigate the specialized social network in which each node is a squash player, and there is a link between any pair of players who have played a match before.

The source data for our study was all US Squash singles matches recorded over a recent multi-year period. The number of players in this network was 26,503 and the number of matches was 240,446. The average number of matches played per player was 18.4 and the maximum was 210 (by Gabriel Bassil of Brooklyn). Like virtually all large-scale social networks, the squash network is sparse, meaning that the number of matches actually played was only a tiny fraction of those possible—less than 7 hundredths of 1 percent. It was also the case that a small number of the most active players account for a disproportionate fraction of the total matches; in network science parlance, the distribution of the number of matches across players is heavy-tailed.

To understand the global shape or structure of our network, we need to examine the *connected components*, which are the islands of connectivity.



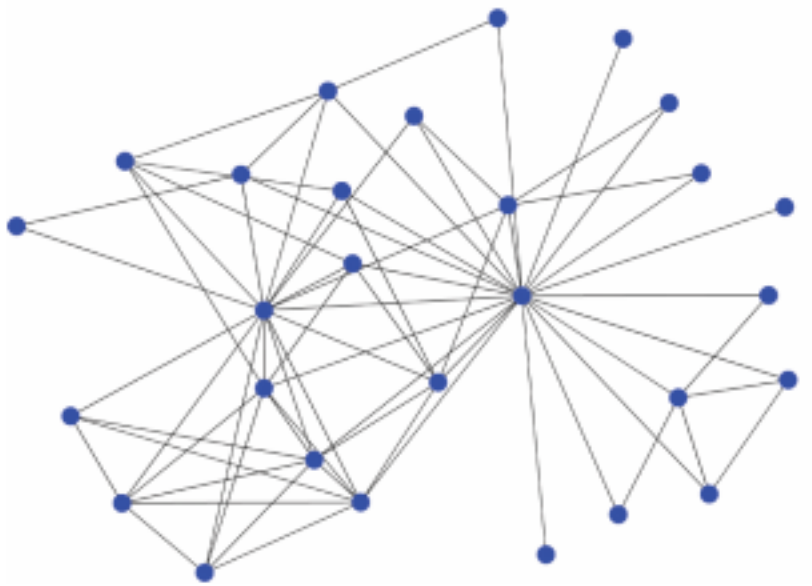
Visualization of the "mainland" of the US Squash network.

Let's consider two players as living in the same island if there is any chain of matches that connects them. So if Alice played Bob, and Bob played Charlie, and Charlie played Dana, then Alice and Dana are in the same connected component (or "island") by virtue of this chain, even if they have never played each other.

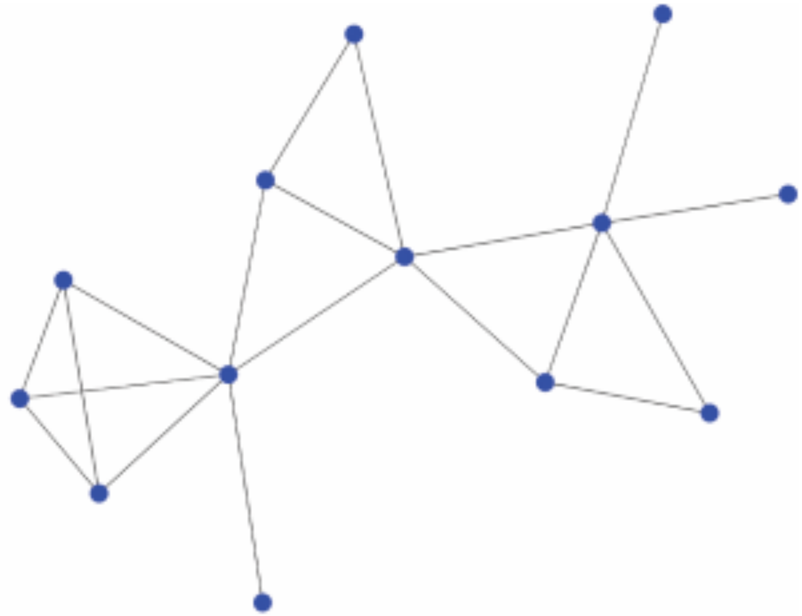
Network science predicts that in any real social network, there should be a giant component—a mainland which contains the vast majority of the population—along with an archipelago of much smaller islands with no links to the mainland. This was the case with our data. The largest component of the squash network contained almost 99% of the players. Intuitively it's hard for two large components to coexist: all it takes is one match between a player from each island and the two merge to become one.

What about the 1% of players in the archipelago, which consists of 77 additional components? What do these tiny islands look like? Unlike Facebook, playing squash requires physical proximity, so it is not surprising that many of the tiny components had a strongly geographic flavor. For instance, the second largest component had only twenty-eight players, all of whom live in Raleigh, NC, while the third largest consisted exclusively of players in San Antonio. Many of the other small components were lonely, isolated pairs of players who had only played each other. We encourage them to play more squash and join the giant component.

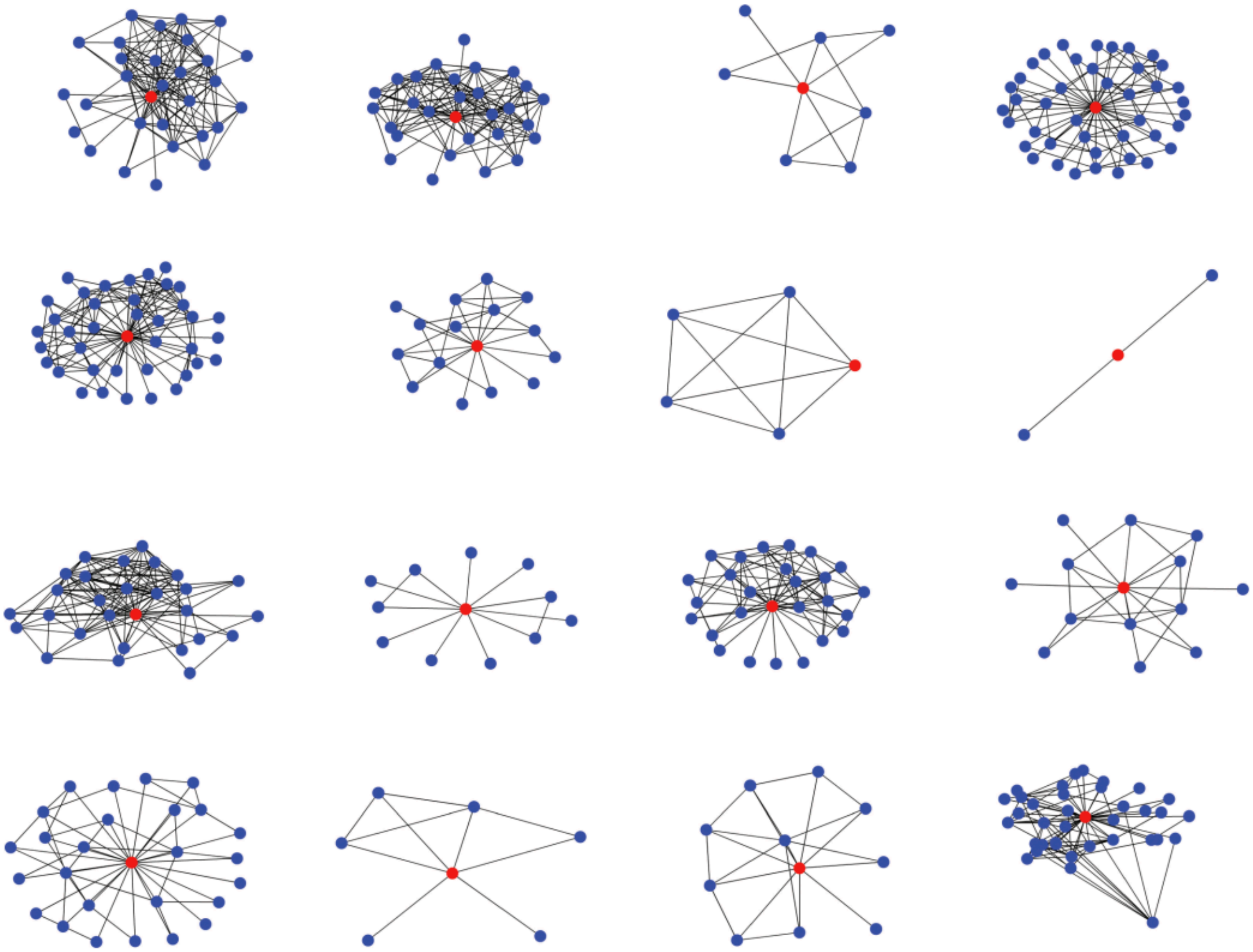
Not all the players in the giant component are



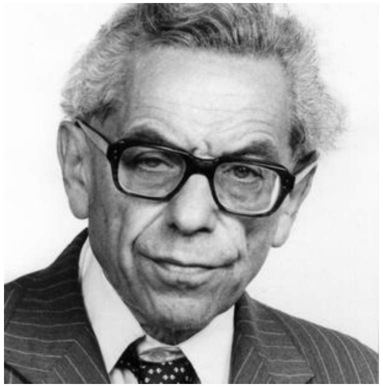
Raleigh NC

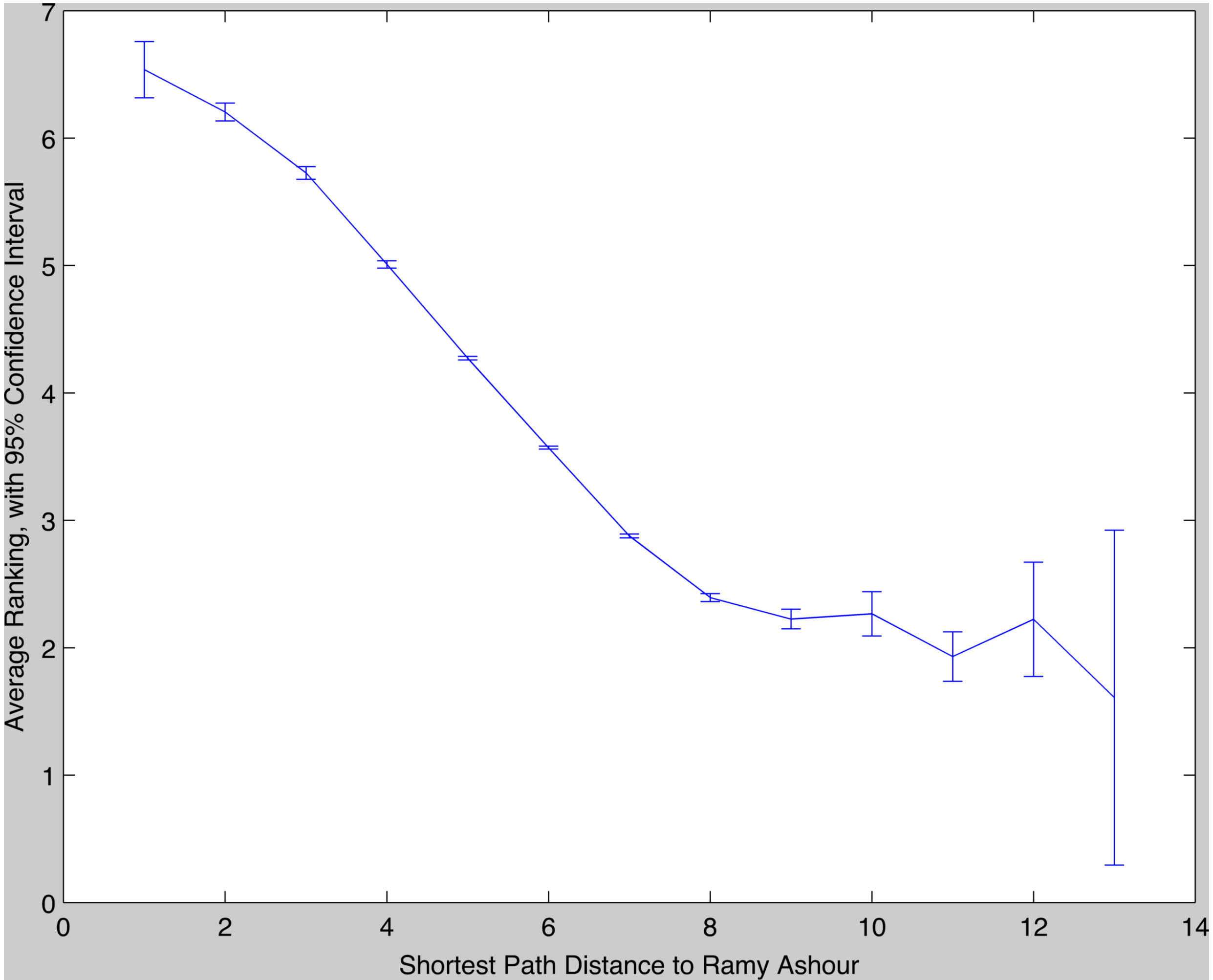


San Antonio TX



Sampling of “Ego” Networks

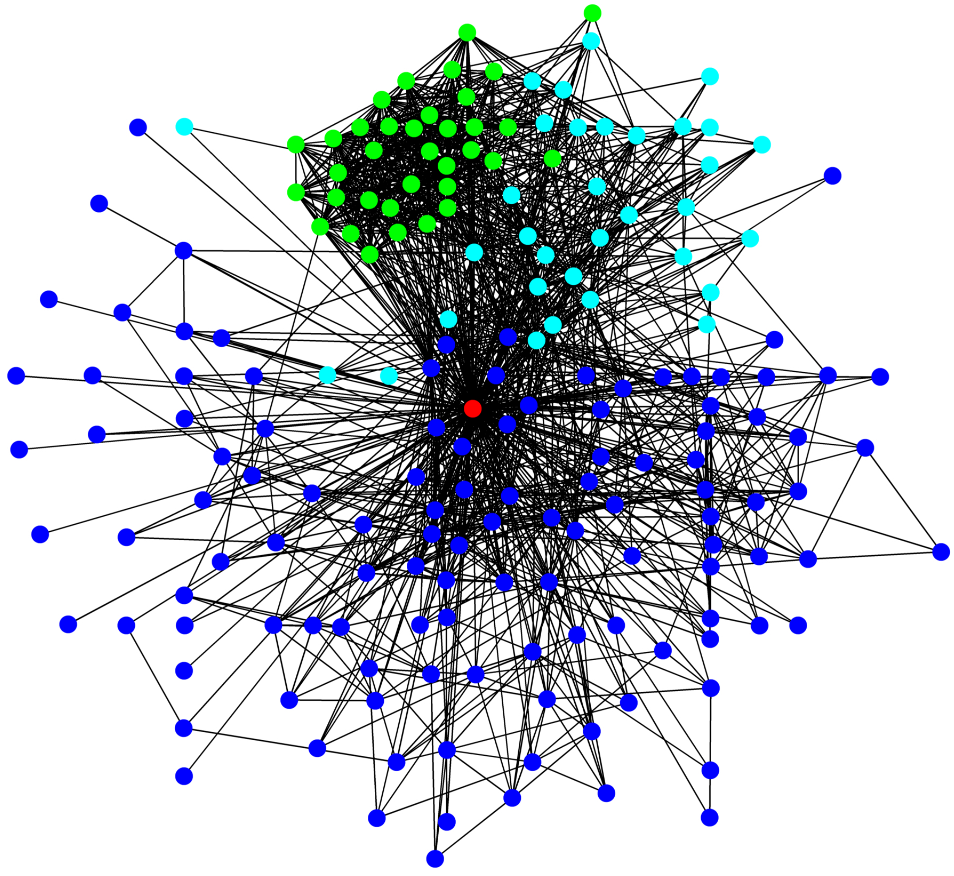




Average Ranking, with 95% Confidence Interval
Shortest Path Distance to Ramy Ashour
0
1
2
3
4
5
6
7
0
2
4
6
8
10
12
14

Measures of Vertex “Importance”

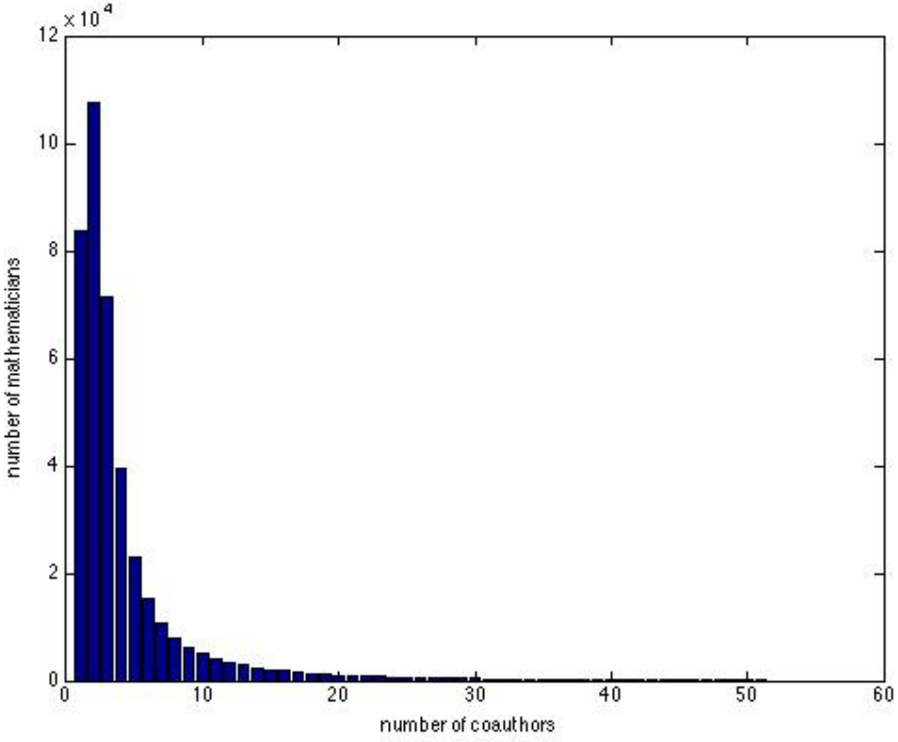
- Exogenous: famous/accomplished/influential/etc individuals
- “Hubs”: high-degree individuals
- Centrality: individuals in the “middle” of the network
- How are these related?



most central squash player, local network

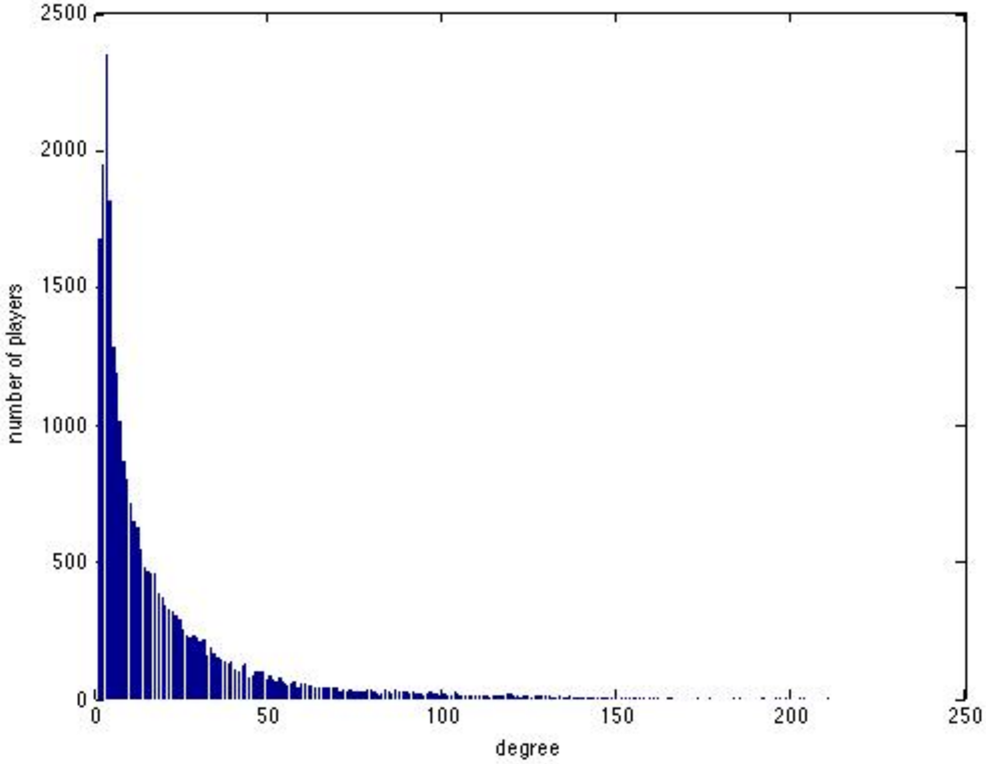
Math Collaboration Degree Distribution

- x axis: number of neighbors/coauthors (degree)
- y axis: number of mathematicians with that degree



Squash Network Degree Distribution

- x axis: number of opponents (degree)
- y axis: number of players with that degree

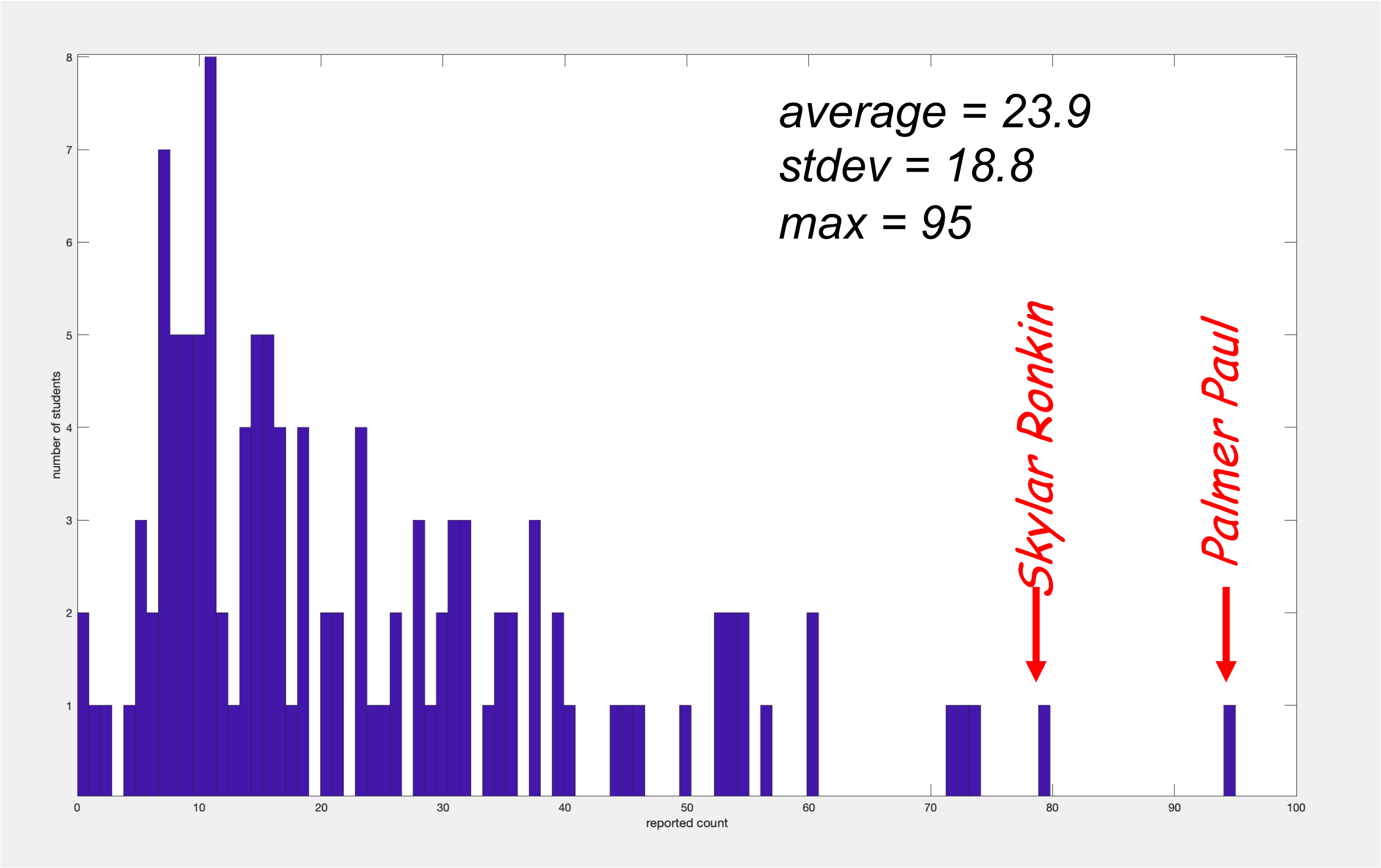


Illustrating the Concepts

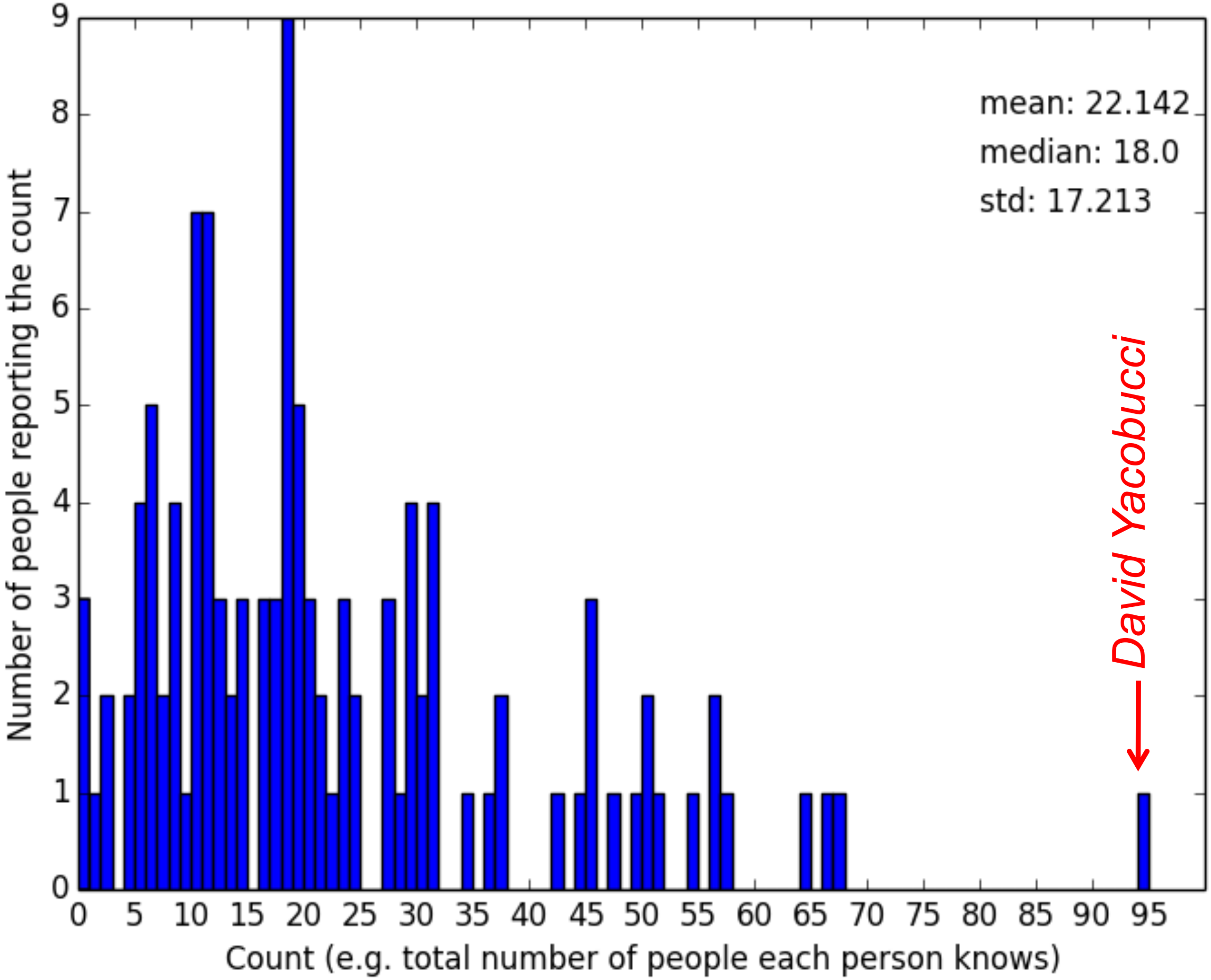
- Example: "real-world" acquaintanceship networks
  - vertices: people in the world
  - links: have met in person and know last names
  - hard to measure
  - let's examine the results of our own *last-names exercise*

Algazi, Alvarez, Alpern, Ametrano, Andrews, Aran, Arnstein, Ashford, Bailey Ballout, Bamberger, Baptista, Barr, Barrows, Baskerville, Bassiri, Bell, Bokgese, Brandao, Bravo, Brooke, Brightman, Billy, Blau, Bohen, Bohn, Borsuk, Brendle, Butler, Calle, Cantwell, Carrell, Chinlund, Cirker, Cohen, Collas, Couch, Callegher, Calcaterra, Cook, Carey, Cassell, Chen, Chung, Clarke, Cohn, Carton, Crowley, Curbelo, Dellamanna, Diaz, Dirar, Duncan, Dagostino, Delakas, Dillon, Donaghey, Daly, Dawson, Edery, Ellis, Elliott, Eastman, Easton, Famous, Fermin, Fialco, Finklestein, Farber, Falkin, Feinman, Friedman, Gardner, Gelpi, Glascock, Grandfield, Greenbaum Greenwood, Gruber, Garil, Goff, Gladwell, Greenup, Gannon, Ganshaw, Garcia, Gennis, Gerard, Gericke, Gilbert, Glassman, Glazer, Gomendio, Gonzalez, Greenstein, Guglielmo, Gurman, Haberkorn, Hoskins, Hussein, Hamm, Hardwick, Harrell, Hauptman, Hawkins, Henderson, Hayman, Hibara, Hehmann, Herbst, Hedges, Hogan, Hoffman, Horowitz, Hsu, Huber, Ikiz, Jaroschy, Johann, Jacobs, Jara, Johnson, Kassel, Keegan, Kuroda, Kavanau, Keller, Kevill, Kiew, Kimbrough, Kline, Kossoff, Kotzitzky, Kahn, Kiesler, Kosser, Korte, Leibowitz, Lin, Liu, Lowrance, Lundh, Laux, Leifer, Leung, Levine, Leiw, Lockwood, Logrono, Lohnes, Lowet, Laber, Leonardi, Marten, McLean, Michaels, Miranda, Moy, Marin, Muir, Murphy, Marodon, Matos, Mendoza, Muraki, Neck, Needham, Noboa, Null, O'Flynn, O'Neill, Orlowski, Perkins, Pieper, Pierre, Pons, Pruska, Paulino, Popper, Potter, Purpura, Palma, Perez, Portocarrero, Punwasi, Rader, Rankin, Ray, Reyes, Richardson, Ritter, Roos, Rose, Rosenfeld, Roth, Rutherford, Rustin, Ramos, Regan, Reisman, Renkert, Roberts, Rowan, Rene, Rosario, Rothbart, Saperstein, Schoenbrod, Schwed, Sears, Statosky, Sutphen, Sheehy, Silverton, Silverman, Silverstein, Sklar, Slotkin, Speros, Stollman, Sadowski, Schles, Shapiro, Sigdel, Snow, Spencer, Steinkol, Stewart, Stires, Stopnik, Stonehill, Tayss, Tilney, Temple, Torfield, Townsend, Trimpin, Turchin, Villa, Vasillov, Voda, Waring, Weber, Weinstein, Wang, Wegimont, Weed, Weishaus.

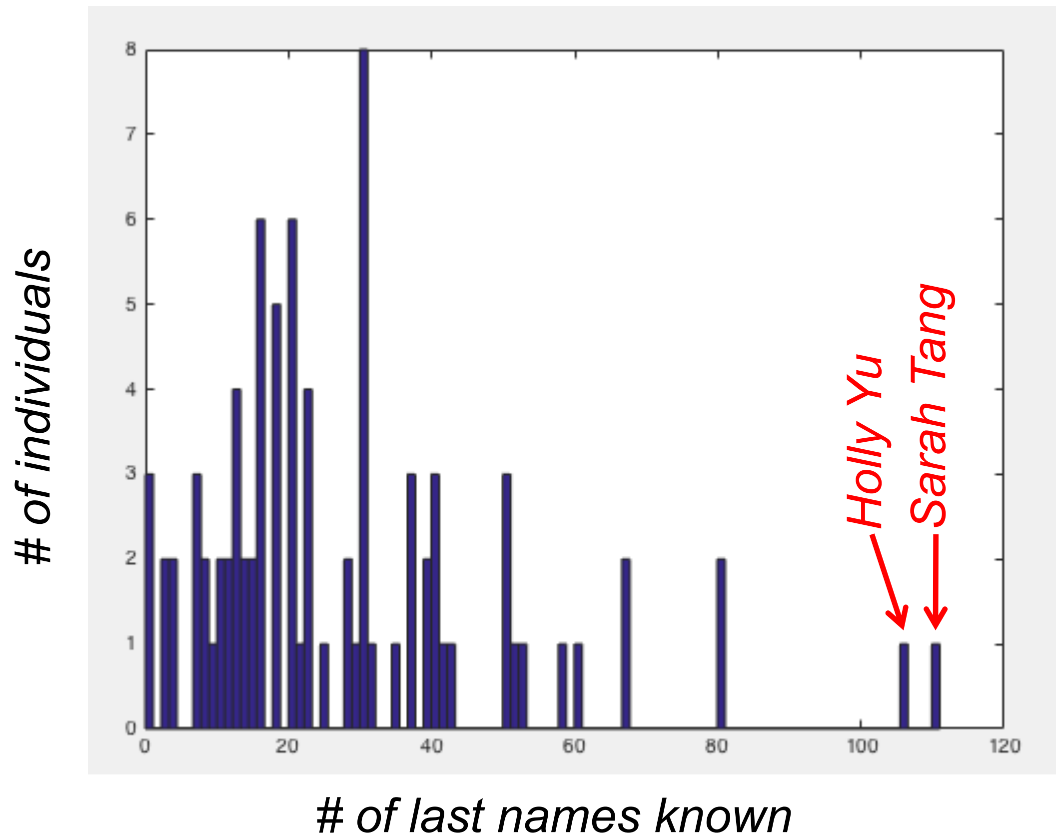
# of individuals



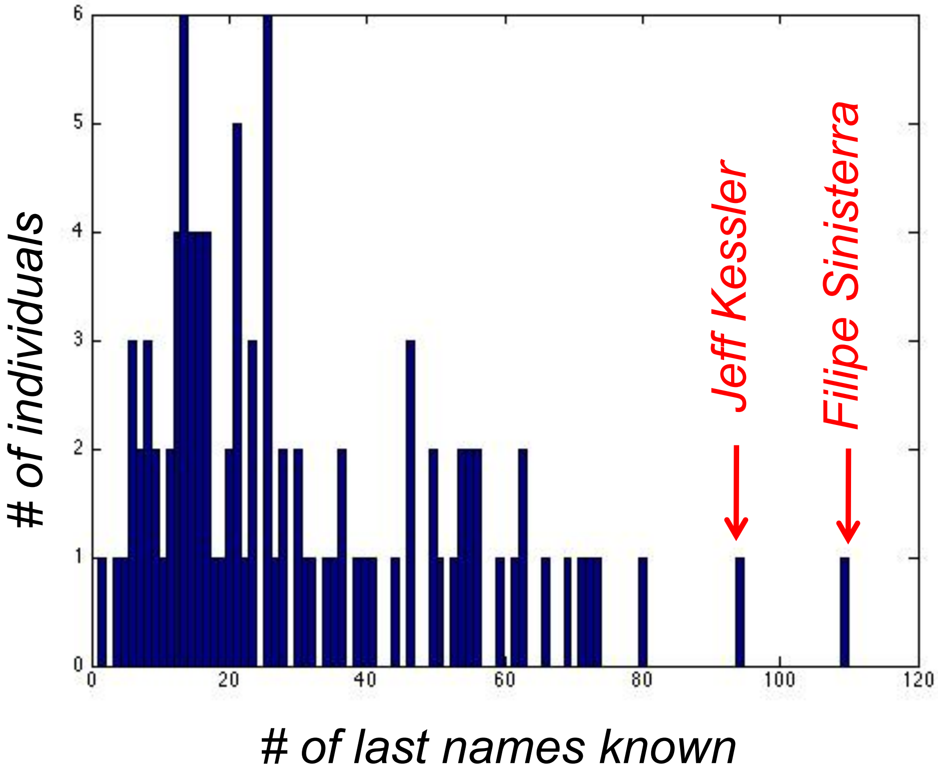
reported count



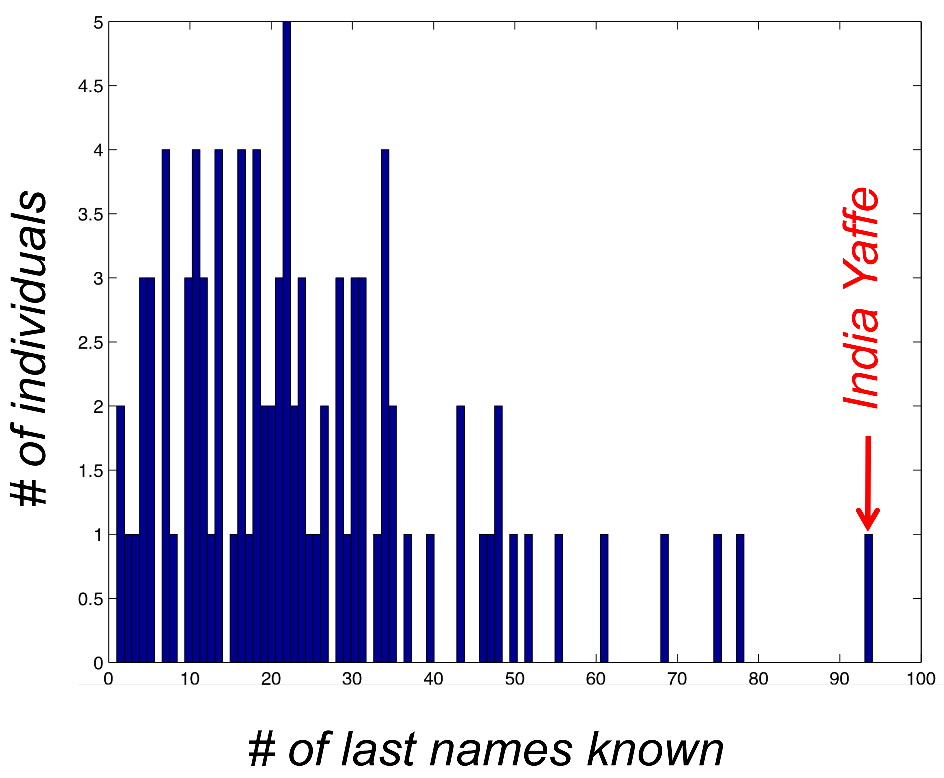
mean: 22.142
median: 18.0
std: 17.213
David Yacobucci
Number of people reporting the count
Count (e.g. total number of people each person knows)
0
1
2
3
4
5
6
7
8
9
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95



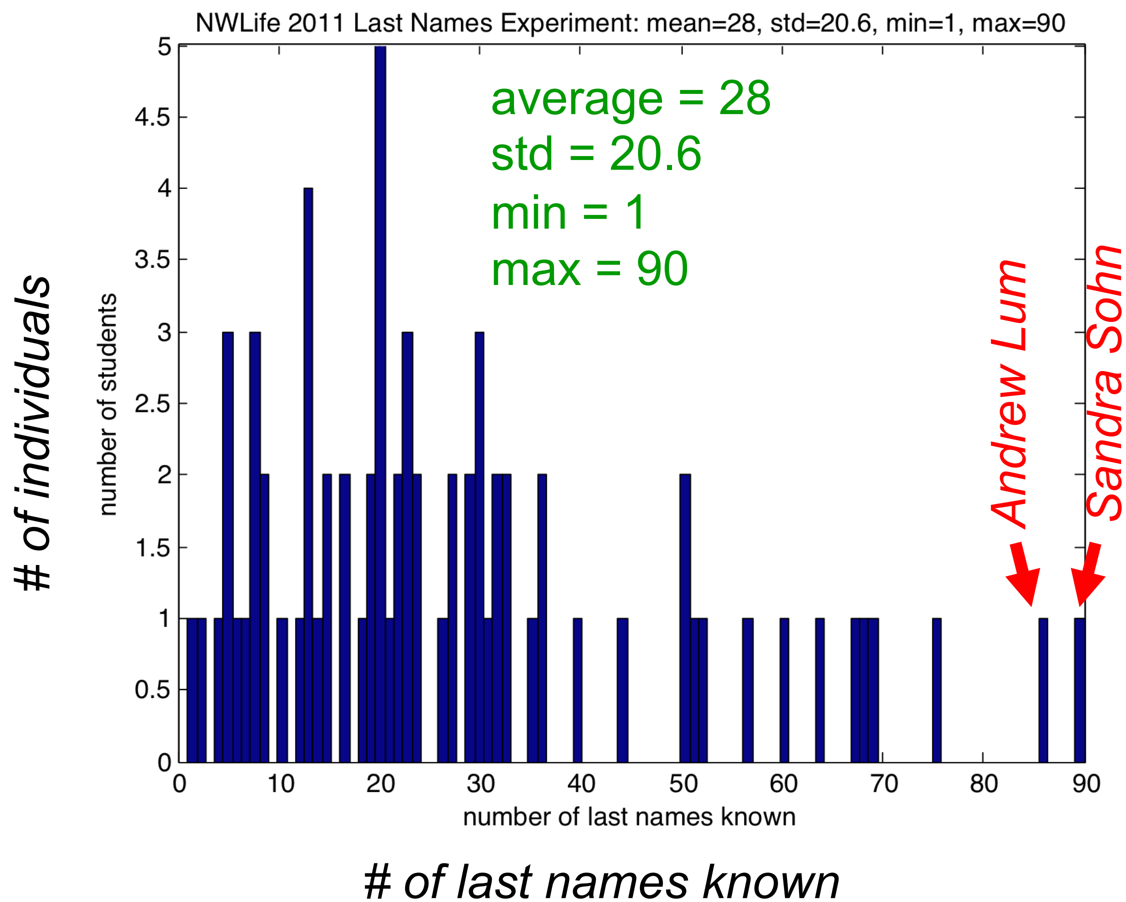
average = 27.9
std = 21.8
min = 0
max = 111



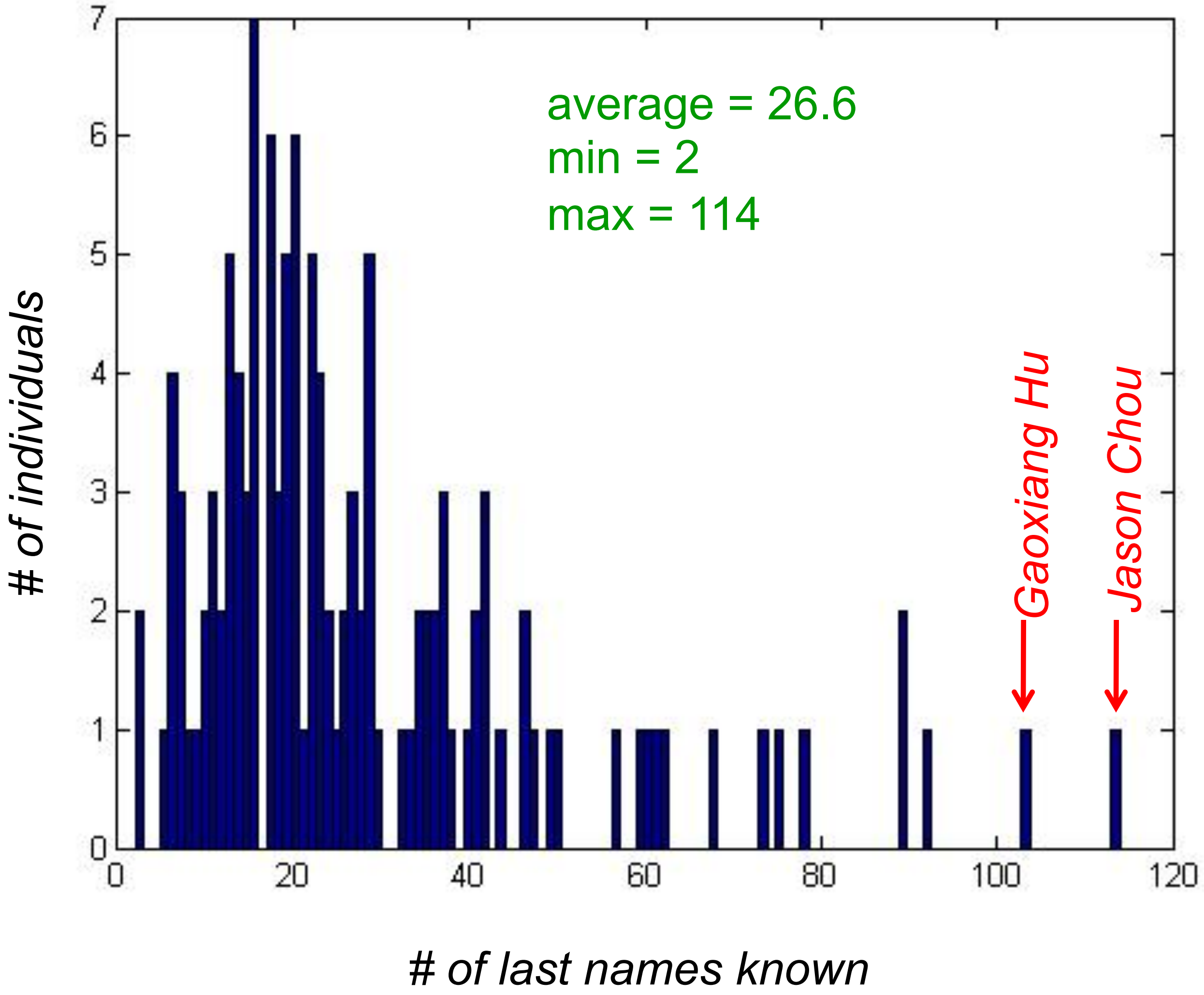
average = 30.2
std = 22.1
min = 1
max = 110



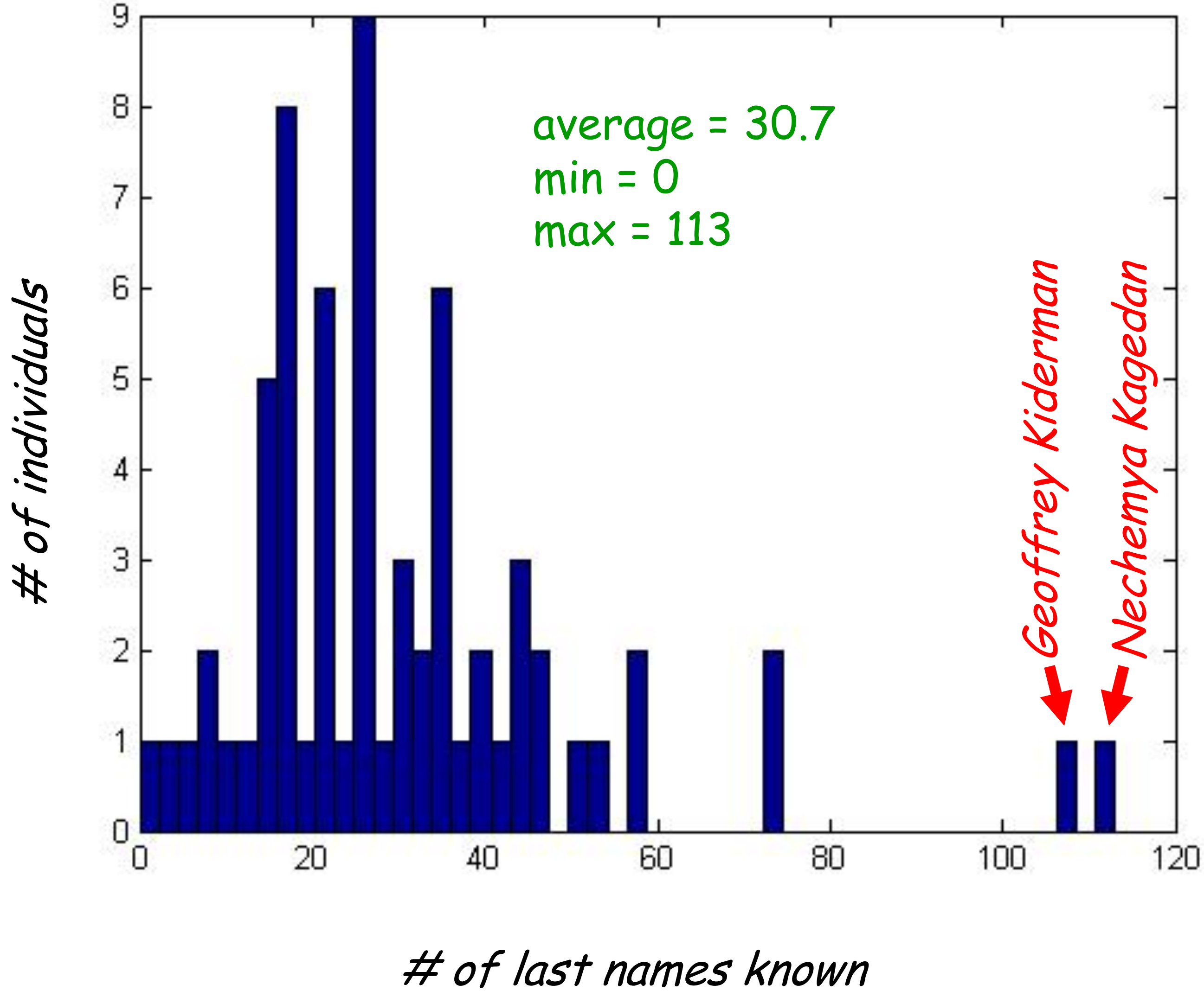
average = 24.6
std = 17.7
min = 1
max = 94



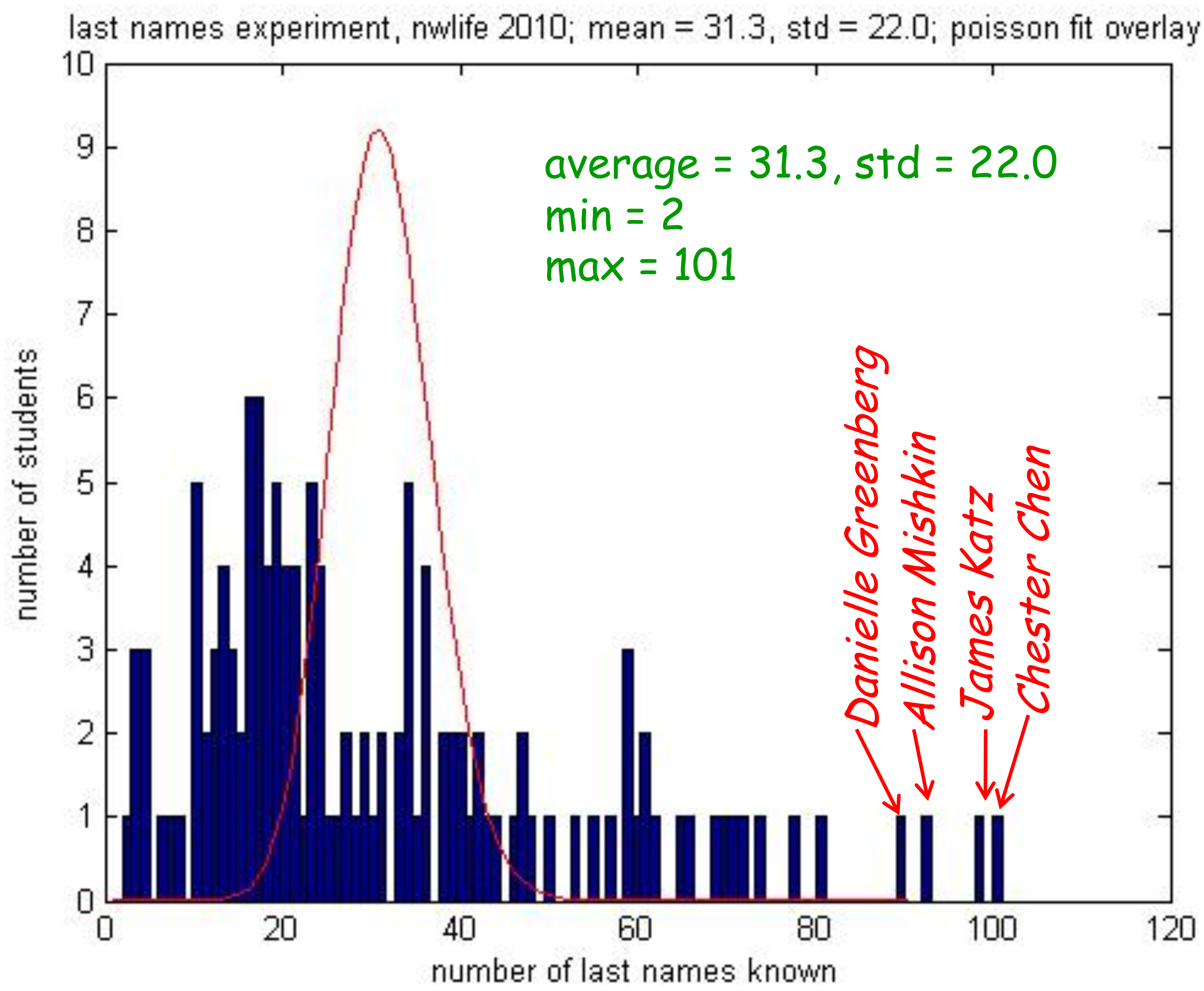
NWLife 2011 Last Names Experiment: mean=28, std=20.6, min=1, max=90
average = 28
std = 20.6
min = 1
max = 90
Andrew Lum
Sandra Sohn
number of students
number of last names known
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
10
20
30
40
50
60
70
80
90
# of individuals
# of last names known

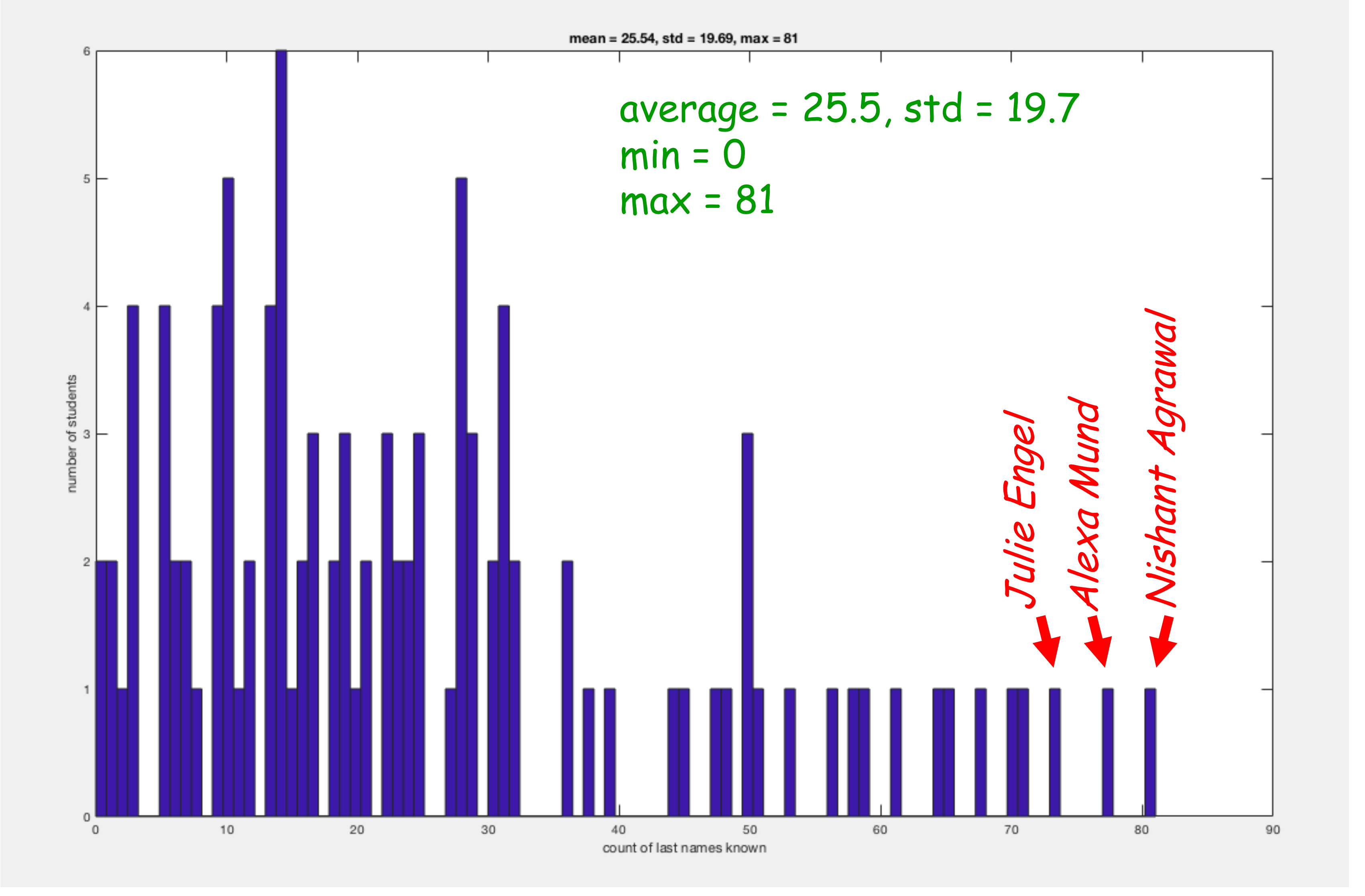


average = 26.6
min = 2
max = 114
Gaoxiang Hu
Jason Chou
# of individuals
# of last names known
0
1
2
3
4
5
6
7
0
20
40
60
80
100
120



average = 30.7
min = 0
max = 113
Geoffrey Kiderman
Nechemya Kagedan
# of individuals
# of last names known
0
1
2
3
4
5
6
7
8
9
0
20
40
60
80
100
120


last names experiment, nwlife 2010; mean = 31.3, std = 22.0; poisson fit overlay
average = 31.3, std = 22.0
min = 2
max = 101
Danielle Greenberg
Allison Mishkin
James Katz
Chester Chen
number of students
number of last names known
0
1
2
3
4
5
6
7
8
9
10
0
20
40
60
80
100
120


mean = 25.54, std = 19.69, max = 81
average = 25.5, std = 19.7
min = 0
max = 81
Julie Engel
Alexa Mund
Nishant Agrawal
number of students
count of last names known
0
1
2
3
4
5
6
10
20
30
40
50
60
70
80
90

Structure, Dynamics, and Formation

Network *Structure (Statics)*

- Emphasize purely *structural* properties
  - size, diameter, connectivity, degree distribution, etc.
  - may examine statistics across many networks
  - will also use the term *topology* to refer to structure
- Structure can reveal:
  - community
  - “important” vertices, centrality, etc.
  - robustness and vulnerabilities
  - can also impose *constraints* on dynamics
- Less emphasis on what actually occurs *on* network
  - web pages are linked… but people surf the web
  - buyers and sellers exchange goods and cash
  - friends are connected… but have specific interactions

Network *Dynamics*

- Emphasis on what *happens* on networks
- Examples:
  - spread of disease/meme/fad in a social network
  - computation of a proper coloring
  - computation in the brain
  - spread of wealth in an economic network
- Statics and dynamics often closely linked
  - rate of disease spread (dynamic) depends critically on network connectivity (static)
  - distribution of wealth depends on network topology
- Dynamics of *transmission* most often studied
- What about dynamics with self-interest, deliberation, rationality?

Network *Formation*

- Why does a particular structure emerge?
- Plausible processes for network formation?
- Generally interested in processes that are
  - decentralized
  - distributed
  - limited to local communication and interaction
  - "organic" and growing
  - consistent with (some) measurement
- The Internet versus traditional telephony