Edge Discovery in a Large Social Network

Jean R. S. Blair[†] and Steven B. Horton[‡]

[†] Department of Electrical Engineering and Computer Science
[‡] Department of Mathematical Sciences
United States Military Academy, West Point

January 2011

Outline

Bridge

- Partnership game
- Uses standard deck of 52 cards dealt equally among 4 players
- Players bid using a common protocol to describe their hand
- At the end of the bidding process, one partner pair has established a contract for how many tricks they will attempt to take
- Game points are won based on the number of tricks taken relative to the contract
- For ACBL sanctioned events, winning partnerships receive master points

Our Goal

Use recorded masterpoints to infer relationships among people

American Contract Bridge League (ACBL)

- 160,000+ members in US, Canada, Mexico, and Bermuda
- 3200 sanctioned clubs
- 1100 tournaments annually
- 3,000,000 tables of bridge played annually in clubs
- 300,000 tables of bridge played annually on-line
- Masterpoint records maintained for all its members

Masterpoints

- Are earned through play in ACBL-sanctioned events
- Are cumulative (for life)
- Result in achieving various ranks

Outline

The Data

Raw Data

- $\sim$ 2.3 GB comma-delimited text file
- $\sim$ 52.5 million records (all available data up to June 2009)

Fields in Each Record

1. Player identifier (ID) as a string of digits
2. Sanction as a string, indicating location of event
3. Event code representing the type of event
4. Rating code indicating the level of the tournament/game
5. Masterpoints received in increments of hundredths of points
6. Pigment of the points received
7. Year that the points were received
8. Month that the points were received
9. Day that the points were received

For example: {8483760, ATLANTA, C, 8, 0.73, R, 1995, 11, 28}

Approach

Recall our goal is to use recorded master points to infer relationships among people

Our Example:
{8483760, ATLANTA, C, 8, 0.73, R, 1995, 11, 28}

Initially extract pairs of rows that are the same except for Player ID:
{6234756, ATLANTA, C, 8, 0.50, R, 1995, 11, 28}
{3434354, ATLANTA, C, 8, 0.73, R, 1995, 11, 28}
{8483760, ATLANTA, C, 8, 0.73, R, 1995, 11, 28}
{2838444, ATLANTA, C, 8, 1.02, R, 1995, 11, 28}

Raw Data Statistics: Event Code

Type of Event	# Occurrences
General Club Points	25,592,937
Women's / men's pairs or teams	170,441
Mixed / unmixed pairs or teams	64,320
Master pairs or teams	306,434
Non-master pairs or teams	706,574
Consolation	197,352
Board-a-match or Swiss teams	5,182,711
Knock-out teams	1,709,517
Individual	15,631
Open pairs	15,813,730
Side game (normally in pairs)	496,862
Charity pairs	2,271,036

The Data

Raw Data

- $\sim$ 2.3 GB comma-delimited text file
- $\sim$ 52.5 million records

The Data

Raw Data

- $\sim$ 2.3 GB comma-delimited text file
- $\sim$ 52.5 million records

Filter Process

- Remove the 54 records with values out of range
  - 26 player ID errors
  - 06 year errors
  - 22 month errors
- Filter out non-pair events and non-pair rating types

The Data

Raw Data

- $\sim$ 2.3 GB comma-delimited text file
- $\sim$ 52.5 million records

Filtered Data

- $\sim$ 0.8 GB comma-delimited text file
- $\sim$ 18.5 million records

Filtered Data Statistics: Players and Masterpoints

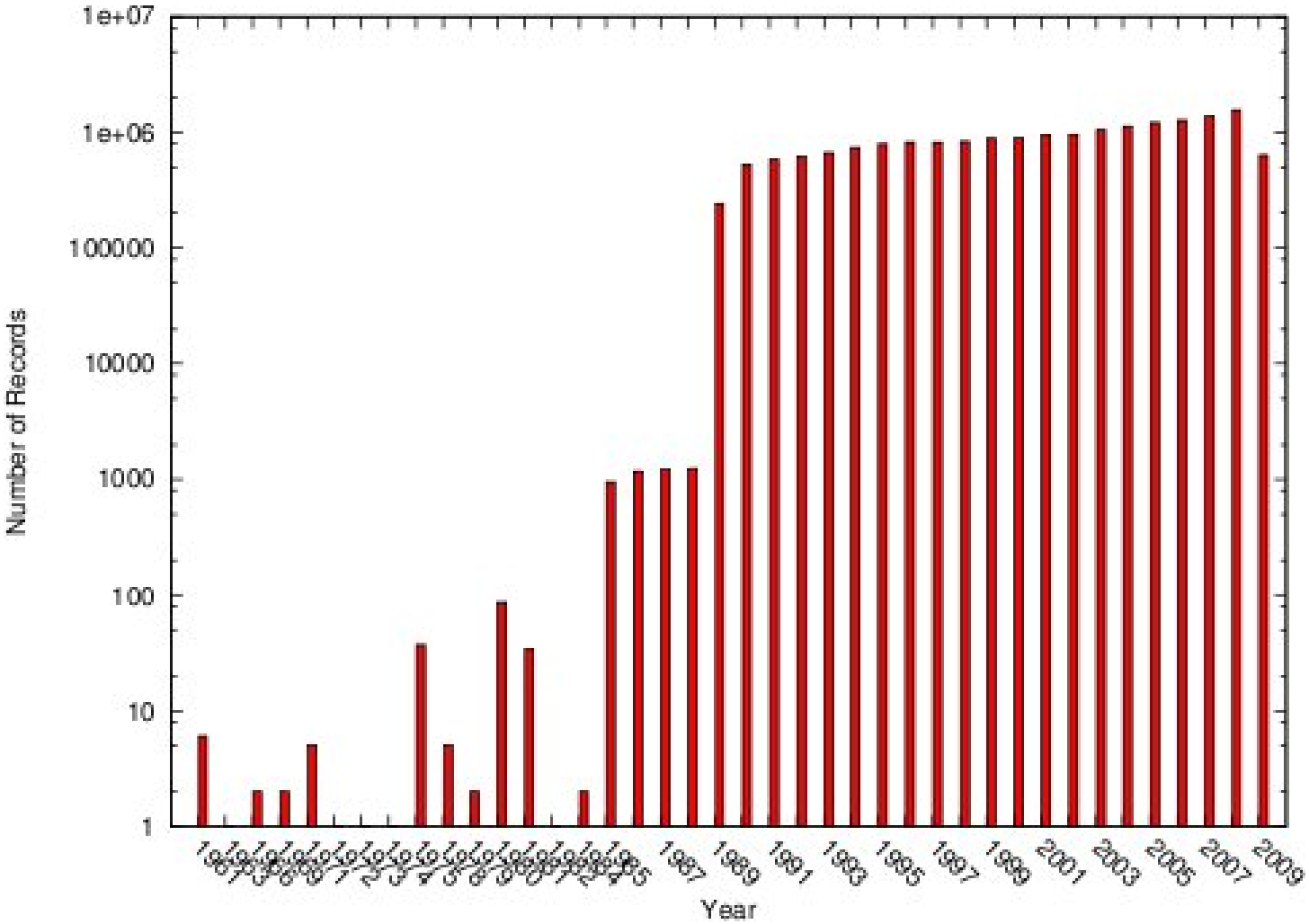
Players

- 281,369 distinct players
- 65.87 records per player, on average
- Number of records / player ranges from 1 to 2,652

Points received

- 1.42185 points / record, on average
- Points received range from 0.01 to 306.00

Filtered Data Statistics: Distribution across Years



Outline

Record Equivalence and Partner Candidate Groups

Definition

A record with ID x is called an *x-record*.

Definition

Two records are said to be *equivalent* if and only if all fields except the ID field have identical values.

Definition

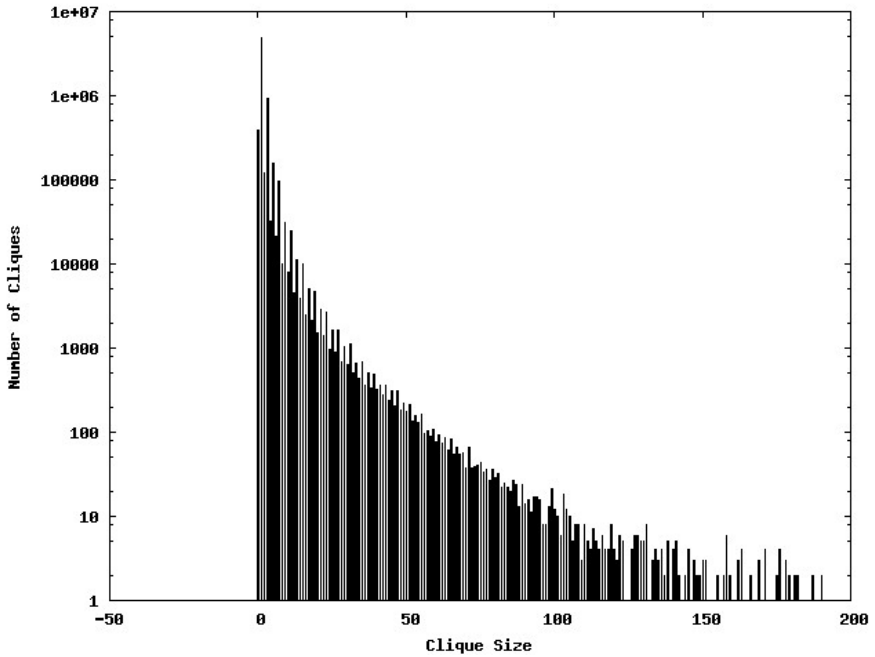
A *partner candidate group* is a maximal set of equivalent records.

Definition

Every partner candidate group of size 2 is a *partner pair*. Every pair of distinct IDs in a partner candidate group of size greater than 2 is a *candidate partner pair*.

Filtered Data Statistics: Partner Candidate Groups

- 6,711,721 partner candidate groups
- Maximum partner candidate group size is 298 (one PCG of this size)
- 4,827,855 partner pairs (partner candidate groups of size 2)



Outline

Inferring Who Played (and won masterpoints) with Whom

Explore different heuristics for inferring partnerships

1. Greedy: Iteratively remove all equivalent instances of known partner pairs, revealing new partner pairs when a partner candidate group is reduced to size 2. Issues include deciding in which order to process known partner pairs.
2. Optimize: Assign partner pairs in order to optimize some metric.
3. Others?

Social Network Questions

1. Does the choice of tournaments to play in impact likelihood of interaction with certain people and/or partnership with certain people?
2. Can you determine characteristics/behaviors of individuals based on partnership data?

Different Graph Models

Consider different graph models defined by:

- Only size-2 partner candidate groups
- Only weight $\geq k$ edges, for some k
- Only records for a given tournament/given month/given state/etc.
- Inducing on vertices within distance k of a given set of players
- Inferred Partner Graphs

Questions about these graphs:

- How do they compare to other social network models?
- Do they exhibit the small-world property?
- What do their connected components look like?

A Little Mathematics

Clustering Coefficients

- If i is a vertex with degree d_i, then its clustering coefficient is $c_i = \frac{2}{d_i(d_i-1)}|e_{jk}|$, or the number of triangles divided by number of possible triangles
- This gives rise to one standard graph clustering measure, $\bar{C} = \frac{1}{n}\sum c_i$, where n is the number of non-infinite c_is and the sum is taken over those terms
- Another common clustering coefficient is $\tilde{C}$ which is the fraction of P_3s that are in a triangle
- Still other things called clustering coefficients are in the literature

A Case Study: The October 1995 Graph

Just One of Many!

Graph G_1 of Initial Partner Pairs

- $|V(G_1)| = 30464$, $|E(G_1)| = 12985$
- $\bar{C} = .0224$, and $\tilde{C} = .0236$

Graph G_2 after (Greedy) Edge Discovery

- $|V(G_2)| = 30464$, $|E(G_2)| = 14413$
- $\bar{C} = .0304$, and $\tilde{C} = .0288$

Table: Number of Components of G_2 with n Vertices

n	1	2	3	4	5	6	7	8	9-137	206	222
	7035	7247	1023	402	155	87	64	38	108	1	1

The Biggest Component of the October 1995 Graph

Properties

- $|V(G)| = 222$, $|E(G)| = 235$ (14 "extra" edges)
- $\bar{C} = .0419$, and $\tilde{C} = .0486$
- $Diameter(G) = 31$
- Mean Shortest Path Length is 12.38
  (Random Connected Graphs of this Size and Order have a Mean Shortest Path Length of About 10)
- This Graph has 6 triangles
  (Random Connected Graphs of this Size and Order have about 0 triangles!)

P-value Idea

Consider a Random Connected Graph on 222 Vertices and 235 Edges (in the spirit of Erdős-Rényi)

- Start with a Random Tree (using Prüfer Codes)
- Add edges between a randomly selected pair of vertices one at a time (keeping the graph simple)
- Intuition: it is quite unlikely that an added edge forms a triangle (less than 1 chance in 100)
- Our October graph has 6 triangles out of only 14 opportunities!
- Partial Details: Average vertex degree is about 2, so typical number of vertices at distance two from a randomly selected vertex is about 2, and $\frac{2}{219} < \frac{1}{100}$

Visualizing Graphs

Leads to Insights

- p-value idea for number of triangles
- Mean Shortest Path Lengths

Outline

Directions for Further Research

- Network Dynamics
- Distance Calculations for Large Graphs
- Effect of Online Play
- Similar Problems?

Thanks

- U.S. Army Research Office
- Elisha Peterson (Johns Hopkins University)