An Architecture for Collecting Longitudinal Social Data

Jeremy Blackburn, Adriana Iamnitchi
University of South Florida

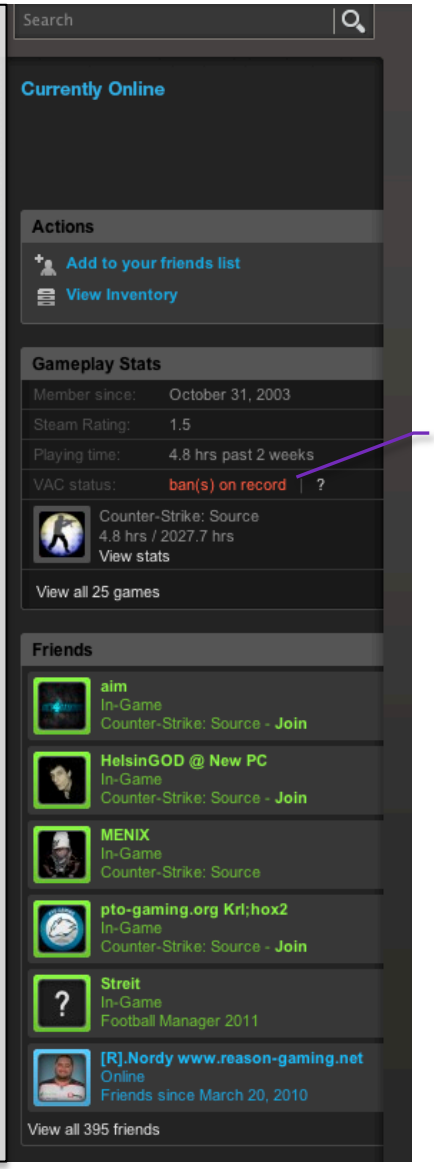
Our Experience: Mining the Steam Community OSN

Steam Community:

- Large online social network for PC gamers
- Built on top of Steam digital delivery platform
- Purchased games permanently tied to account
- Steam account required to create Steam Community profile
  - Steam Community profile not required to play games

Steam Community Profile

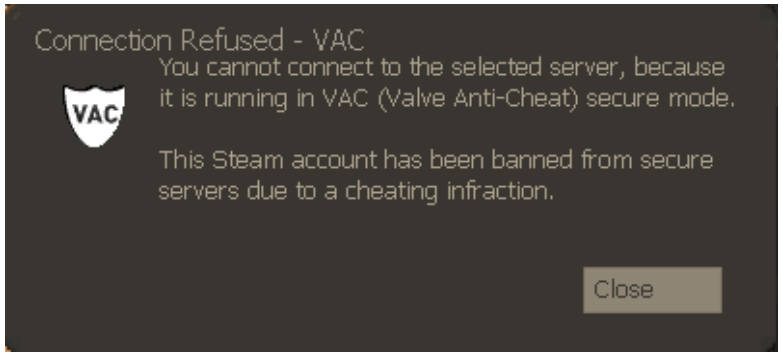
- **Unique SteamID**
- **Friends list**
- **User specified location**
- **Cheating flag (VAC ban)**
- Nickname (mutable)
- Date of account creation
- Screenshots
- Videos
- Comments (“wall posts”)
- Profile information
- Game reviews
- Gameplay ownership/stats
- Virtual goods inventory



Cheating flag

The cheating flag

- Cheating automatically detected via Valve Anti Cheat system
  - Method and timestamp not public
  - Delayed application
- Permanent
- Publicly viewable
  - Even private accounts
- Can't play on VAC secured servers
  - Only applies to the game that was cheated in
- Most servers are VAC secured
  - 4,200 of 4,234 Team Fortress 2 servers
- Cheaters not removed from Steam Community



Why Care About Cheaters in Online Games?

- Cheats (hacks): code that enhances skills (see through walls, automatically aim, move very fast)
- Cheating in online games:
  - serious cost for the gaming industry
  - undisputed unethical behavior
- Cheating affects many aspects of our lives
  - Opportunity to quantify and study spread
- Initial research question: the location of cheaters in the social network: Influential? Clustered together?

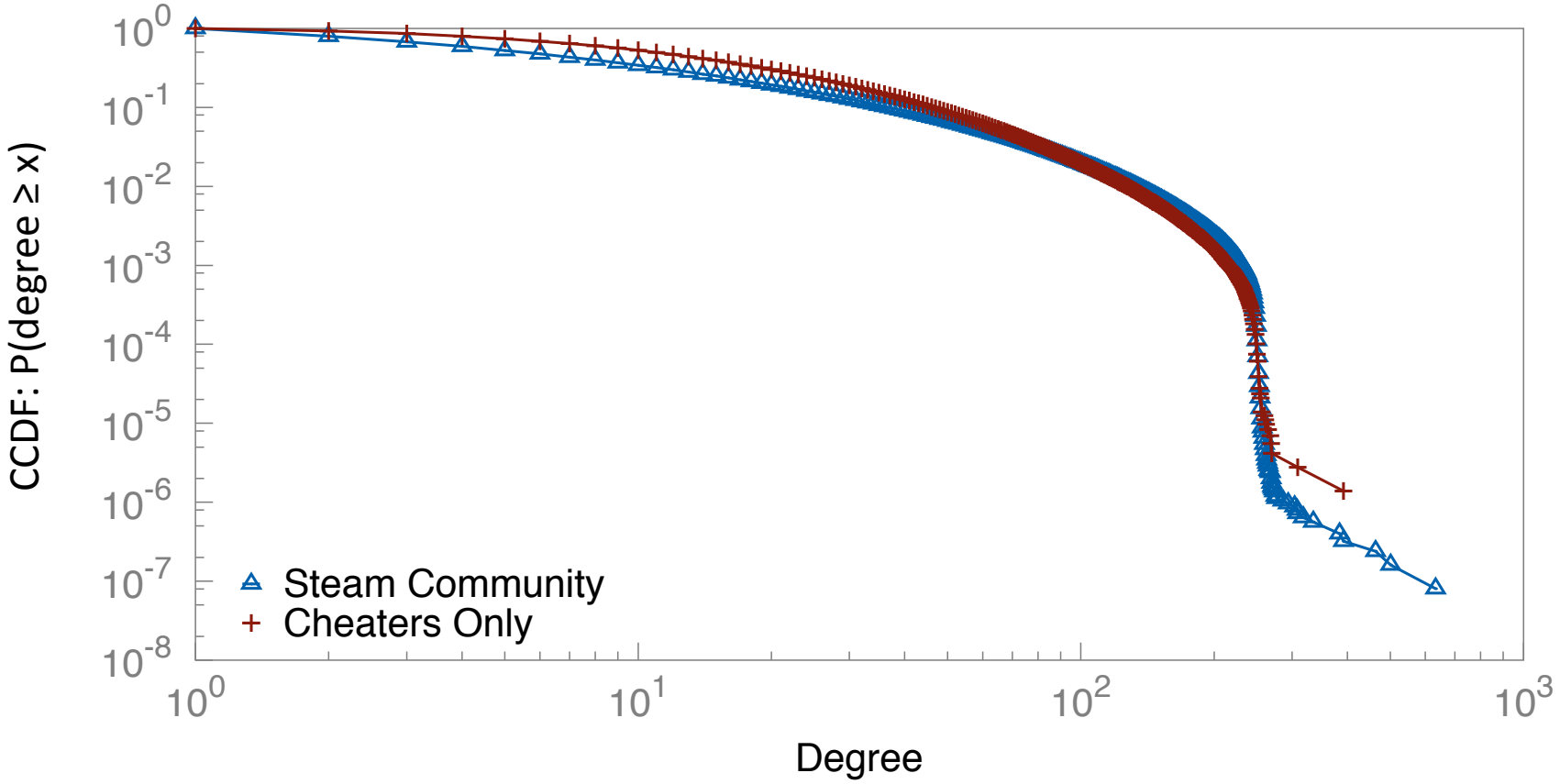
Steam Community data set

Type	Nodes	Edges	Profiles	Public	Private	Friends-only	Location set
All users	12,479,765	88,557,725	10,191,296	9,025,656	313,710	851,930	4,681,829
Cheaters	-	-	720,469	628,025	46,270	46,714	312,354

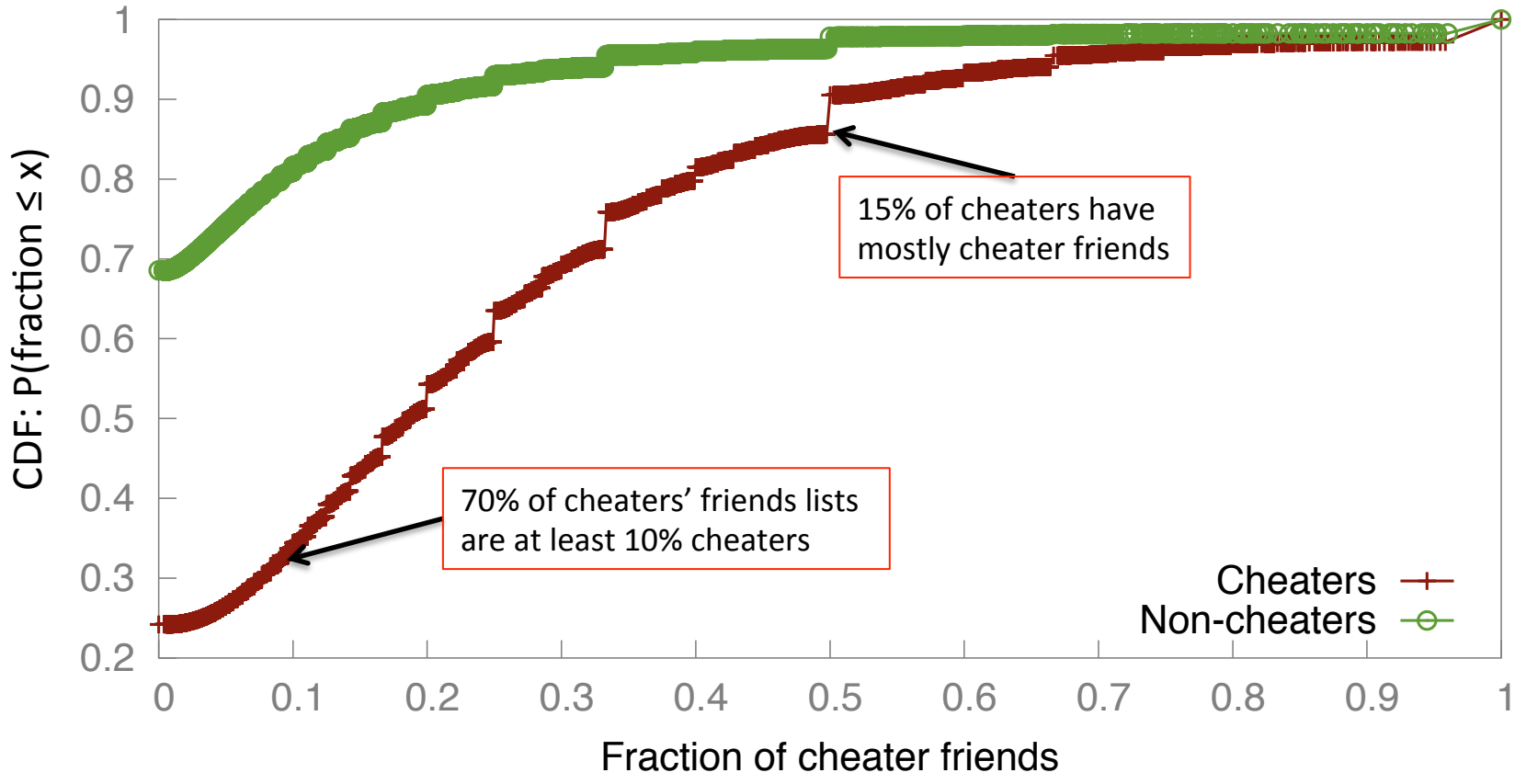
- Data collected March 16 – April 3, 2011
  - Distributed BFS using Amazon EC2
- Cheaters make up 7% of profiles
- 7% of cheaters have private profiles
  - 3% of non-cheaters with private profiles
- Cheaters as likely to be friends-only as private
  - Non-cheaters about 3 times as likely to be friends-only as private

Cheaters more likely to be private than non-cheaters

Are Cheaters Well Embedded in the OSN?

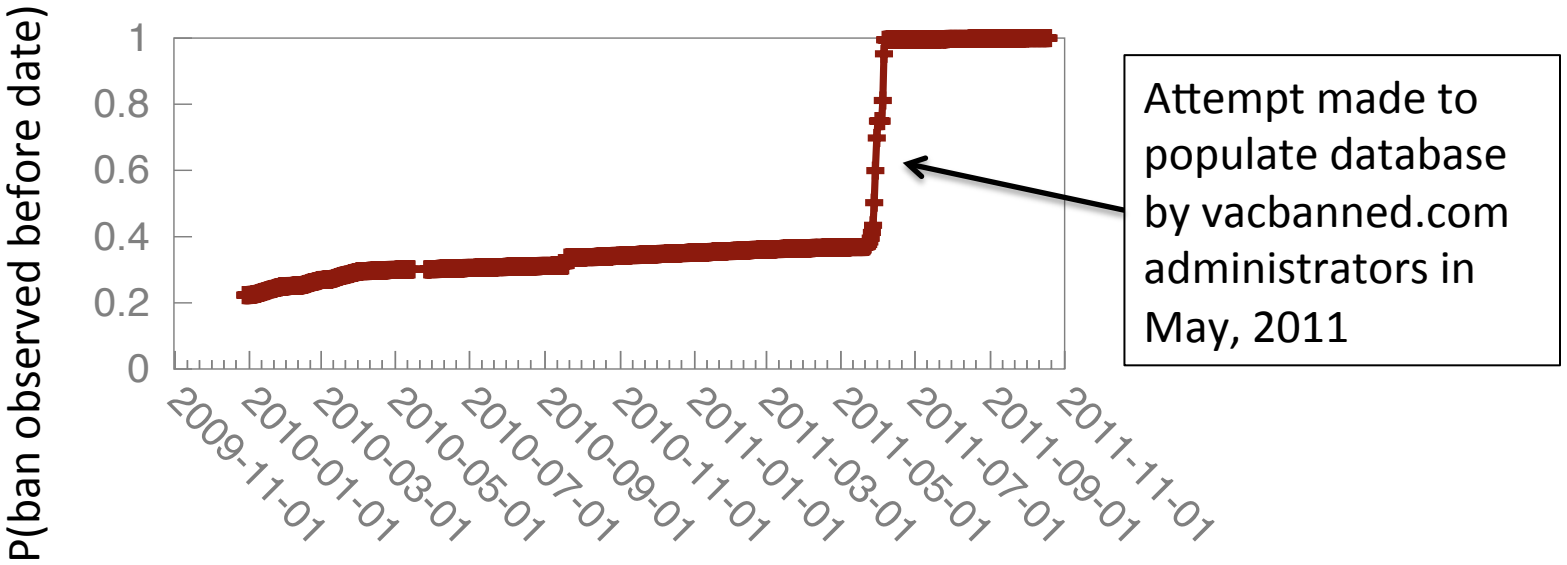


Who is Friends with Cheaters?



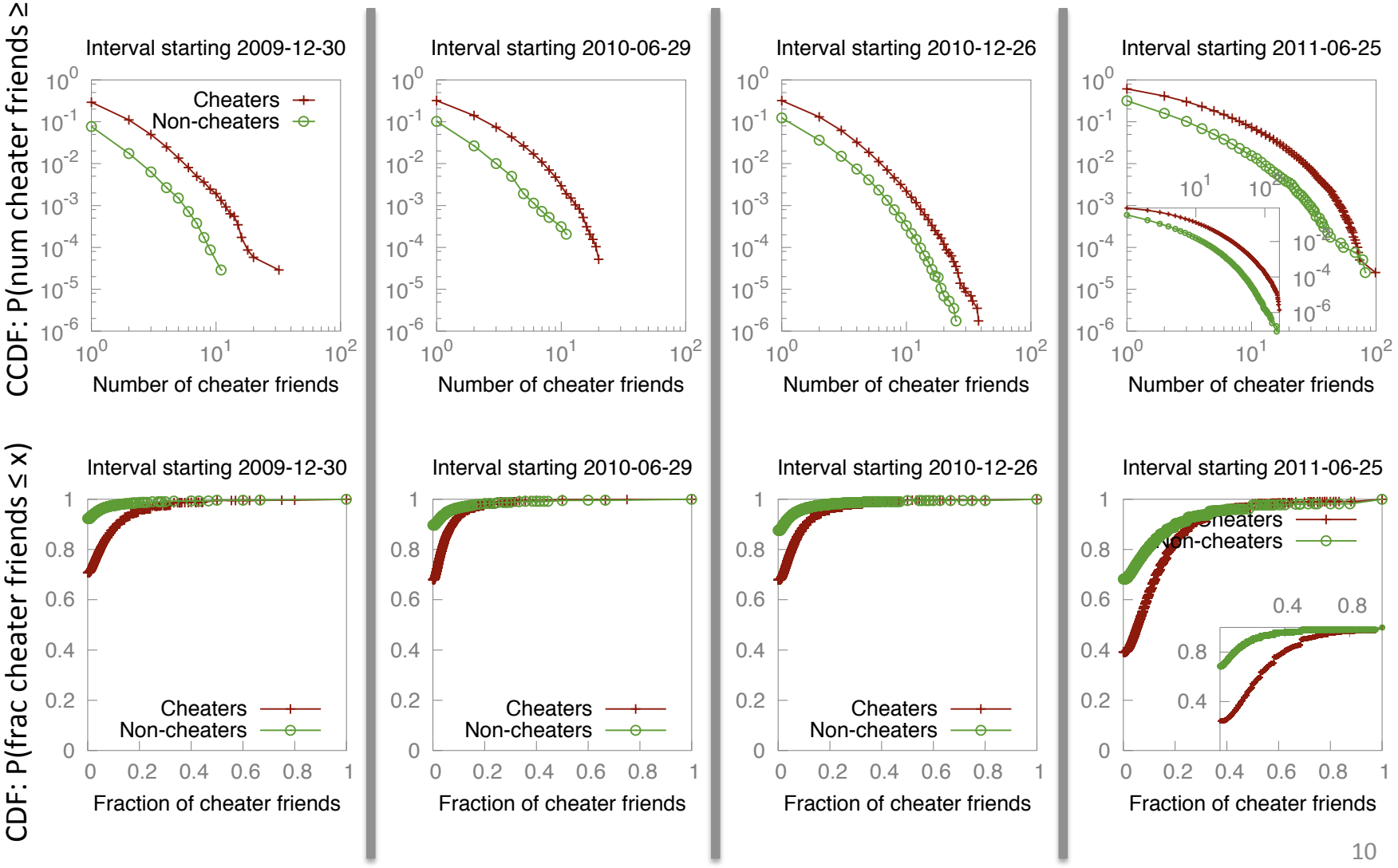
We Needed More Accurate Data…

- 3rd party web site, vacbanned.com, provides historical data on when a VAC ban was first observed
  - Dates must be treated as banned "on or before"



Blackburn et al, "Branded with a Scarlet "C": Cheaters in a Gaming Network", WWW 2012.

Evolution of Cheaters' Social Structure



Challenges with Data Collection

- APIs provided by OSNs are often incomplete, poorly documented or simply not working.
- New functionalities are continuously added, and with them new attributes that prove relevant to data analysis.
- Continuously changing needs both in data and analysis.
- New sources of data might appear.

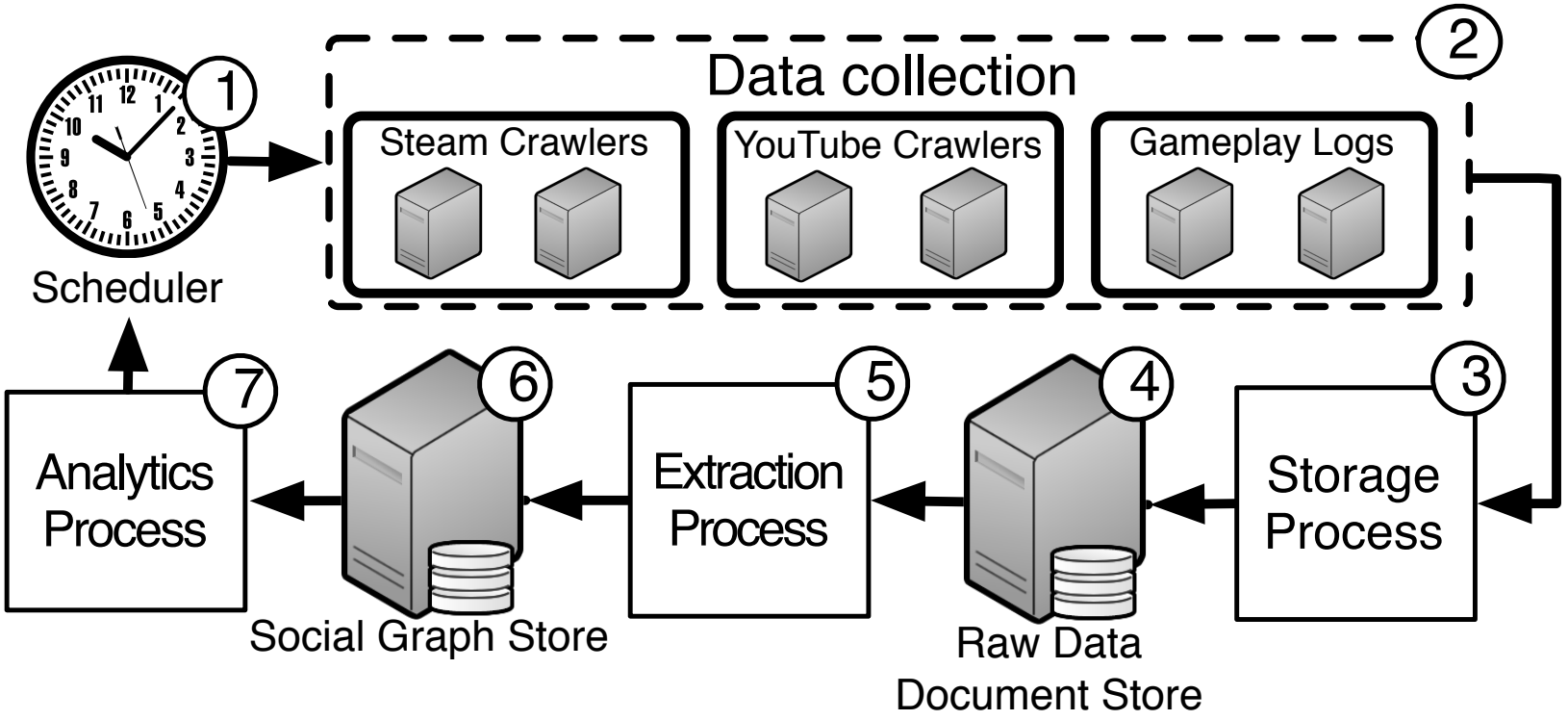
Our Solution (1)

- Decoupling data collection from extraction and analysis
  - successive iterations of data extractions as informed by data analysis results with minimum costs;
  - transparent collection of newly introduced attributes embedded within the data source independent of the data extraction or processing mechanisms;
  - Independent insertion or removal of new sources of data.

Our Solution (2)

- Versioning and provenance
  - At minimum, this requires time stamping of data.
  - Changes in data extraction code requires versioning
- Multigraph, ego-net based model: representing the fusion of multiple sources of social data for a single ego

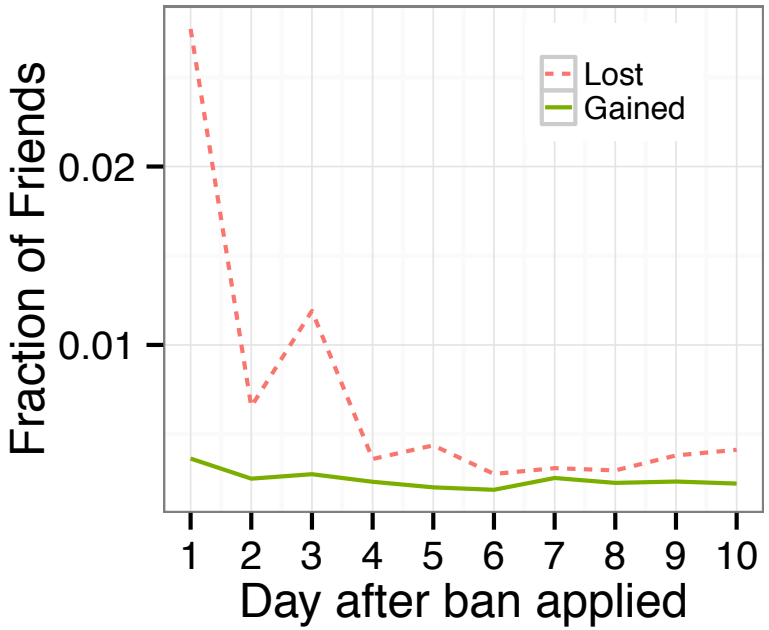
The System


1
2
Data collection
Steam Crawlers
YouTube Crawlers
Gameplay Logs
Scheduler
7
Analytics Process
6
Social Graph Store
5
Extraction Process
4
Raw Data Document Store
3
Storage Process

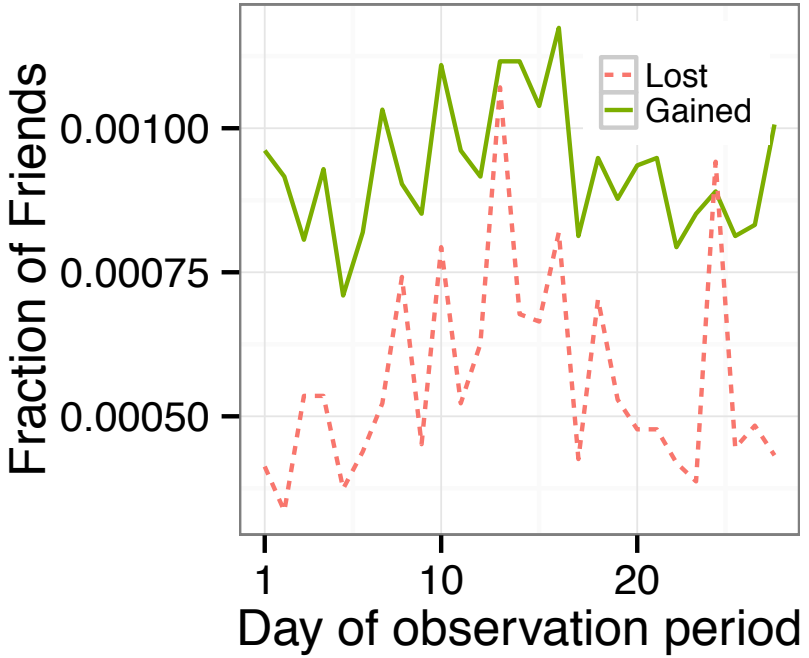
Monitoring Gamers

- Rails application with distributed worker processes
- MongoDB for data storage
  - 9 nodes + 1 application server, each with 32 GB RAM
  - ~155 GB of data (~525M observations)
- At 00:01 users in the system have their ban status queried
  - Deltas stored
- Newly transitioned cheaters have their neighborhood monitored for 10 days
  - Their neighbors added to system too
  - Monitor only if we have at least one observation where the users is not banned

Reaction to Cheating Flag

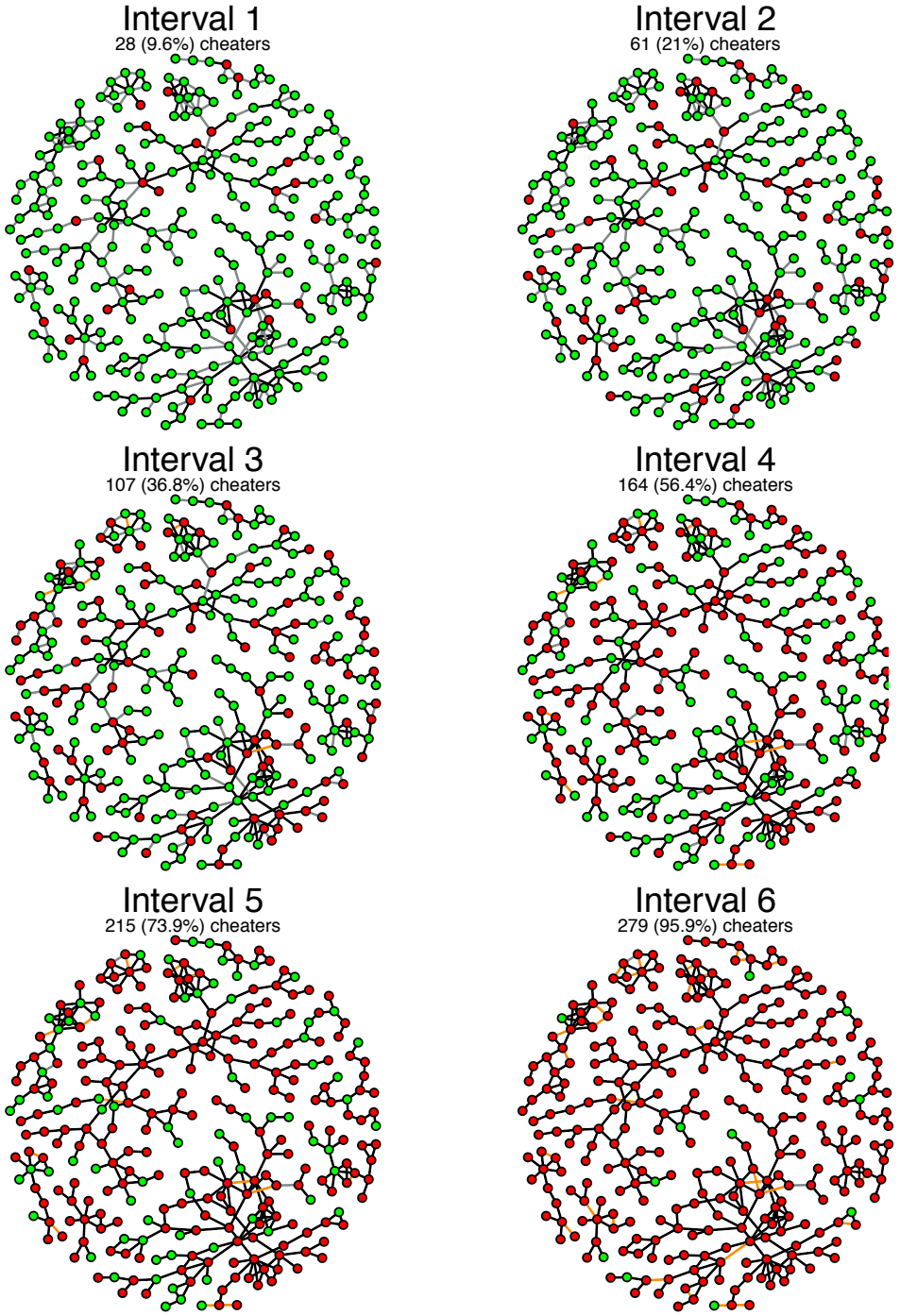


New cheaters



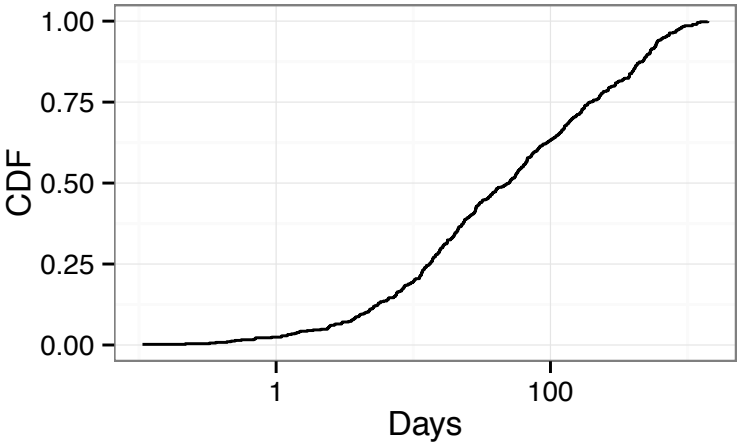
Control group

(both relative to number of friends on first date of observation)



High resolution spread

- 5 day intervals
- 10 largest connected components



Most are friends for a long time before they cheat

Thank you!

An Architecture for Collecting Longitudinal Social Data

Jeremy Blackburn, Adriana Iamnitchi

University of South Florida

http://www.cse.usf.edu/dsg/