

Strange Bedfellows: Community Identification in BitTorrent

David Choffnes

Jordi Duch, Dean Malmgren, Roger Guimerà,
Fabián Bustamante, Luís A. Nunes Amaral

Northwestern University



Privacy in P2P Systems

- Privacy increasingly important, elusive goal
 - As connectivity improves, privacy declines
 - Affects Web browsing, social networks, P2P systems...



- Existing attacks
 - Snoop connections to reveal content
 - Infiltrate system with rogue clients to pollute or spy
 - Interfere with targeted connections
 - ...

Privacy in Swarming Systems

- In P2P swarming, attacks can involve identifying
 - Content that users download
 - Content that users share
 - Who they share it with
- Available countermeasures
 - Encrypt connections
 - Decentralize swarm membership identification
 - Darknets, networks of trust

Are the connections themselves a threat to privacy?

Evaluating Privacy in P2P Systems

- Goal for this work
 - Determine how much information is revealed by *connection patterns* in swarming P2P system
- Simple enough in theory, but...
 - Connections require *simultaneous*, shared interest in content
 - Intimately tied to user behavior, difficult to model
 - Spread of P2P makes empirical connection data difficult to gather
- Ono dataset for connection traces
 - Currently installed by nearly 1,000,000 BitTorrent users
 - Gathers per-connection data (but no info for content)

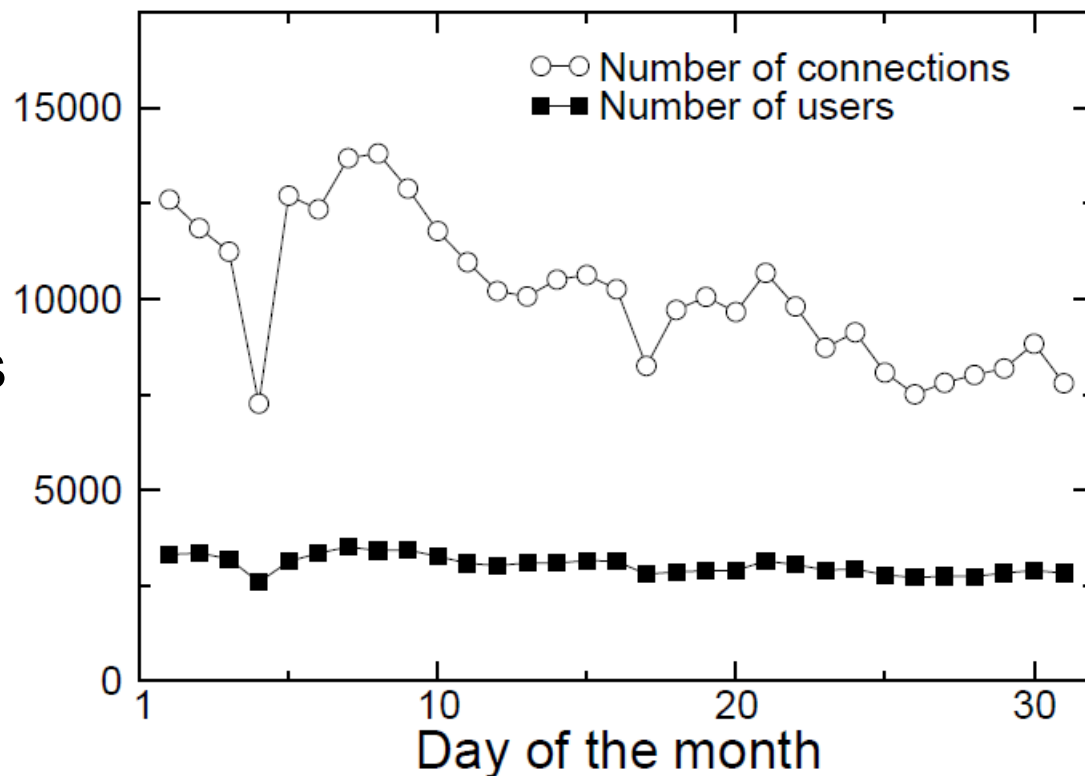
Connection Patterns in BitTorrent

- Is there (global) structure to BT connections?
 - Reasons for
 - People share interest for a variety of content
 - Regularity in time-of-day usage
 - Reasons against
 - Random connections in BitTorrent
 - Difference in transfer rates
 - Selfish behavior (download and depart)
 - Geographic spread of users (time zones)
- Examine structure through graph representation
 - BT hosts are nodes, connections are edges
 - Popular approach: identifying *communities* in the graph

Building a BitTorrent Network Graph

● Dataset

- March 1-31, 2008
- Restricted to Ono users' connections



● Graph representation

- Build weekly graphs (account for weekly patterns)
- Each edge assigned weight according to number of days connected during the week

Communities in BitTorrent

- Do these user connections reveal communities?
 - Can be solved by maximizing modularity

$$\mathcal{M}(\mathcal{P}) = \frac{1}{2L} \sum_{ij} \left[w_{ij} - \frac{s_i s_j}{2L} \right] \delta_{m_i m_j}$$

Given a connection between nodes i and j , how much of their total connection strength is it?

Communities in BitTorrent

- Do these user connections reveal communities?
 - Can be solved by maximizing modularity

$$\mathcal{M}(\mathcal{P}) = \frac{1}{2L} \sum_{ij} \left[w_{ij} - \frac{s_i s_j}{2L} \right] \delta_{m_i m_j}$$

Only count those in the same community

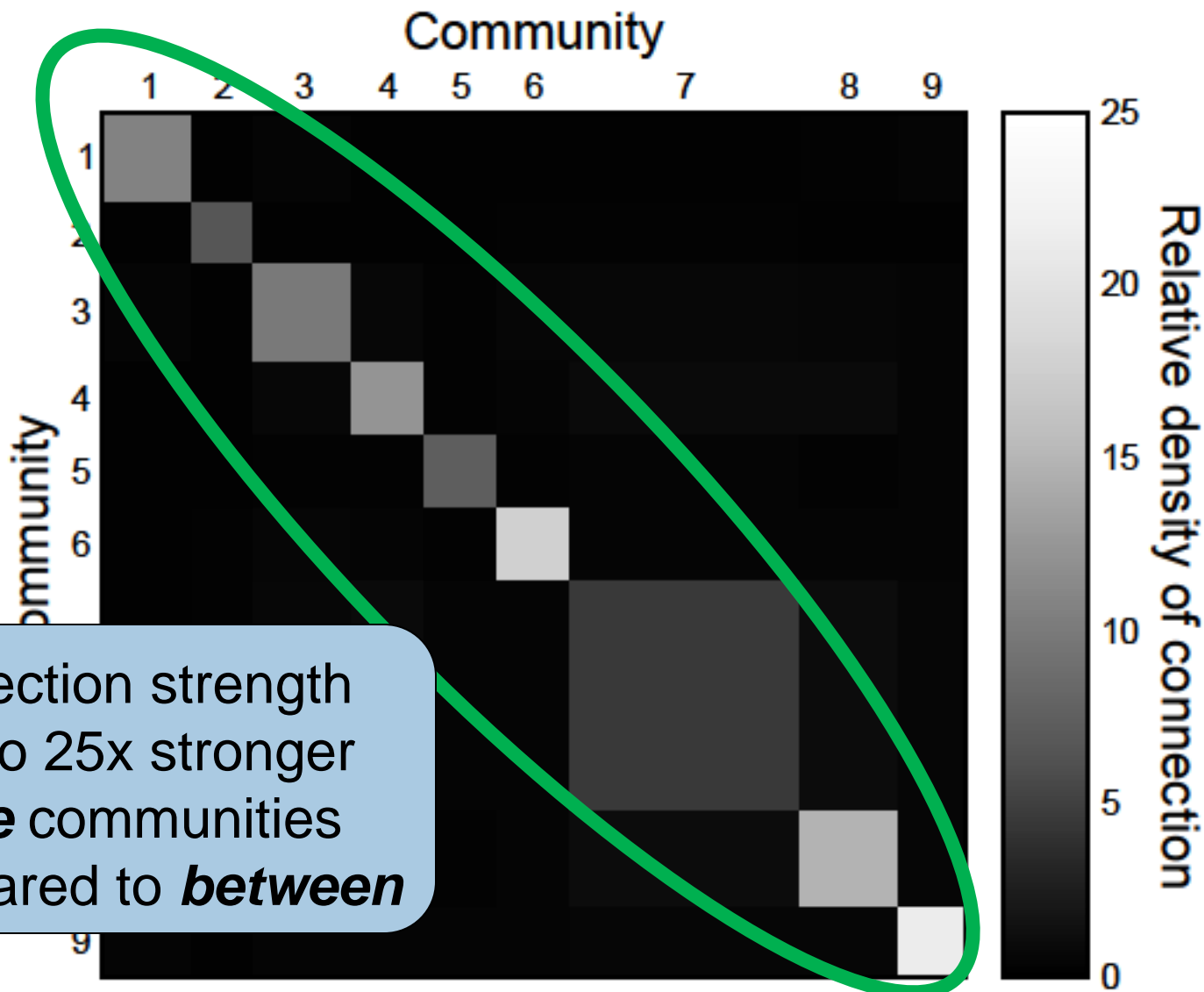
Communities in BitTorrent

- Do these user connections reveal communities?
 - Can be solved by maximizing modularity

$$\mathcal{M}(\mathcal{P}) = \frac{1}{2L} \sum_{ij} \left[w_{ij} - \frac{s_i s_j}{2L} \right] \delta_{m_i m_j}$$

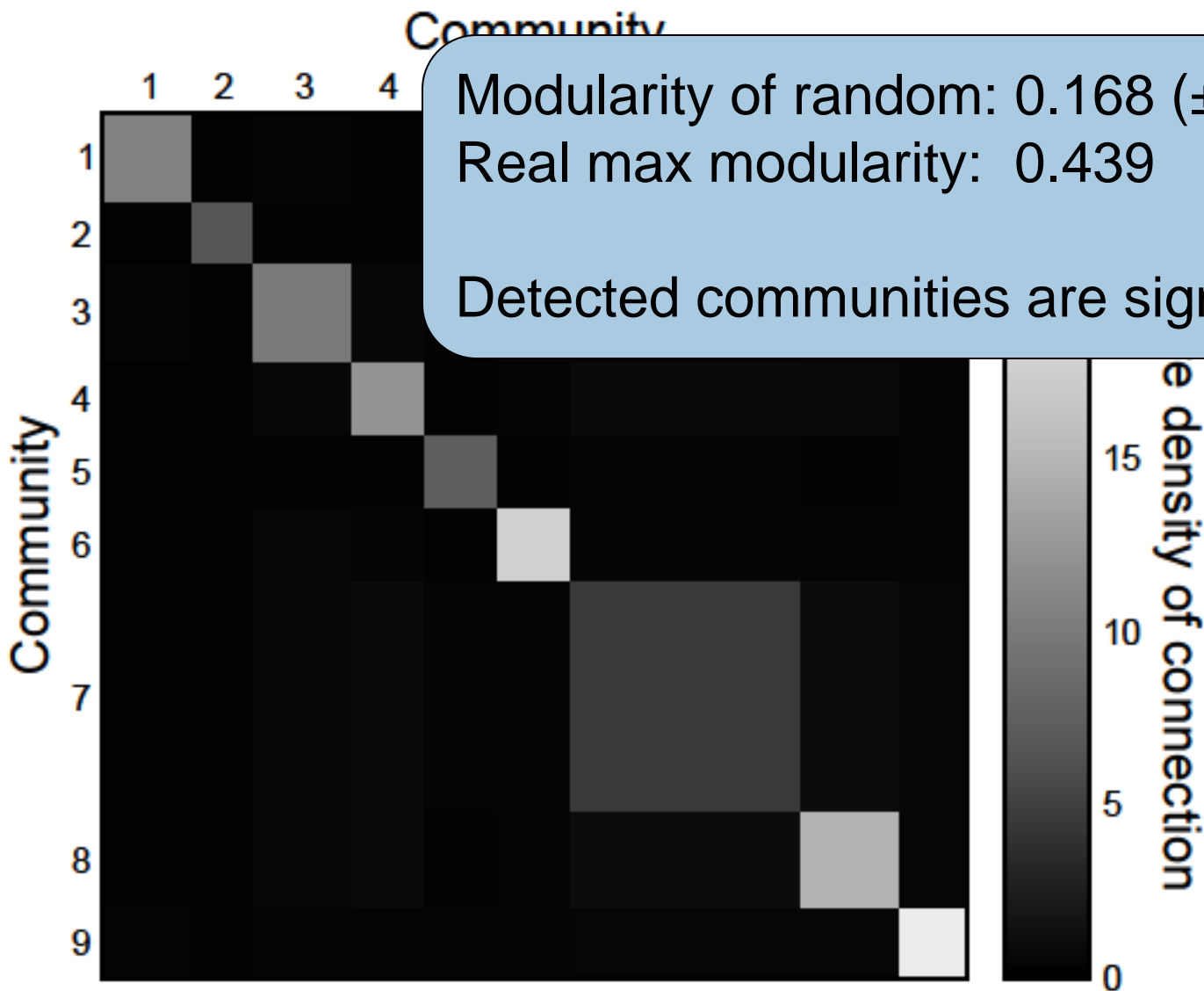
- Determines amount of connection weight *within* communities as opposed to *between* them
- Challenges
 - NP-hard problem
 - Many heuristic techniques
- Extremal optimization
 - Good trade-off between speed and accuracy $O(N^2 \ln N)$
 - Nearly identical to other randomized approaches

Community Identification Results






Connection strength is 5x to 25x stronger *inside* communities compared to *between*

Community Identification Results



Why Do Communities Matter?

- General advantages to communities
 - Allows optimizations based on structure
 - Social networks: Suggest friends 
 - Web browsing: Target advertising 
 - P2P file-sharing systems: Infer content interest 
- Risks of communities in P2P systems
 - Copyright enforcement
 - Censorship
 - **Guilt by association**

Why Do Communities Matter?

- Communities for guilt by association
 - Small numbers of hosts predict behavior of entire group
 - **Not** a legal definition, *per se*
 - Facilitates surveillance, e.g.
- Real world example (McCarthy era)
 - *Alder v. Board of Education of New York* (1952)
 - US law was **upheld**, dissenting opinion:
 - “The present law proceeds on a principle repugnant to our society — guilt by association.[...] Teachers are under constant surveillance...; their utterances are watched for clues to dangerous thoughts.”
 - Justice William O. Douglas

Guilt by association

- Guilt by association in BitTorrent
 1. Identify a community
 2. Identify the content shared by a single member
 3. Infer that all members of the community are doing the same *without monitoring them directly*

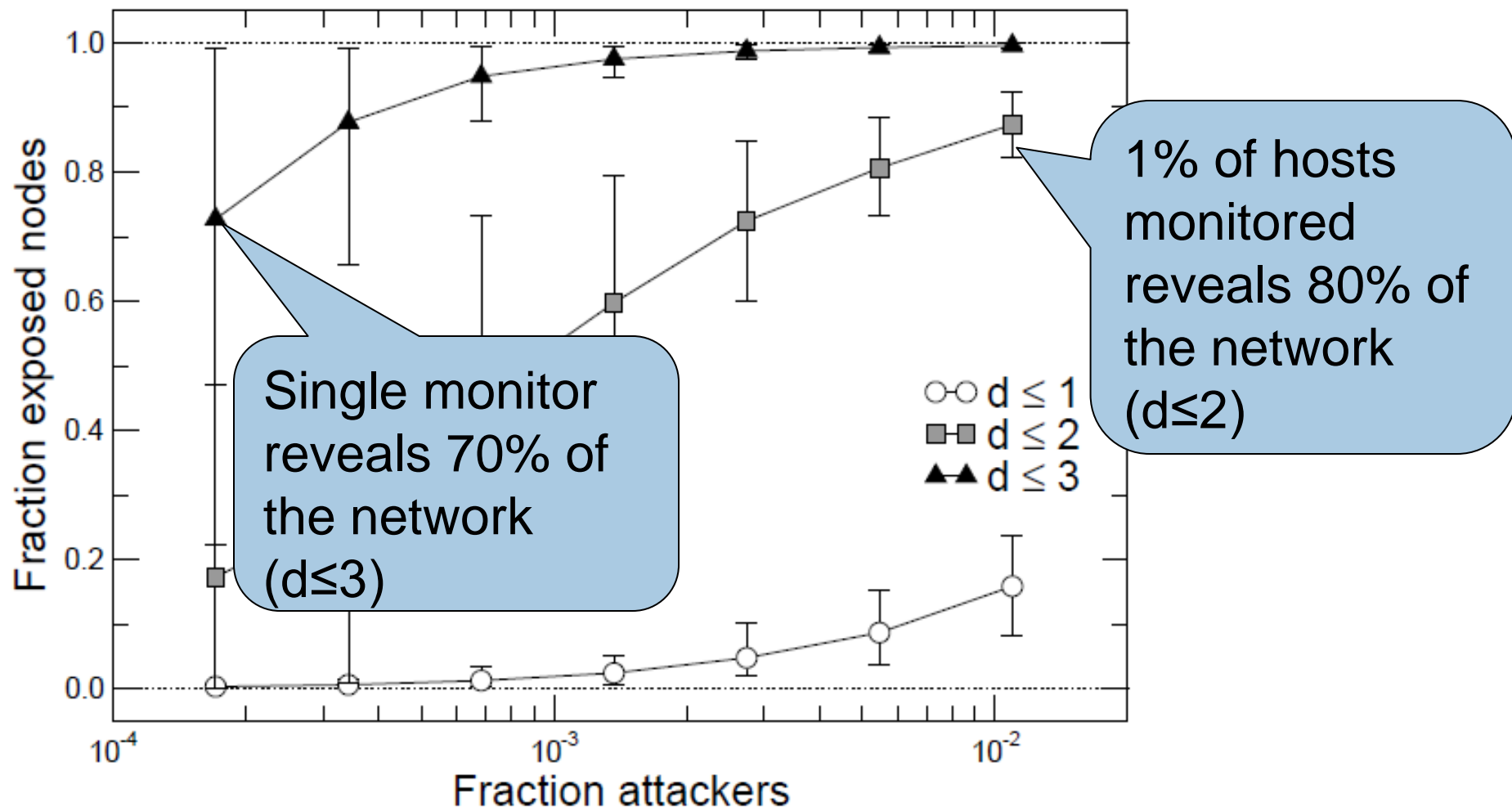
Can this be used to efficiently monitor BT?

Discovering the Connection Graph

- Approaches to building connection graph
 - Trackers: The swarm is the community
 - Difficult with trackerless torrents
 - Limited to per-torrent view
 - Does not reveal connection information
 - Peer Exchange (PEX)
 - Reveals peers' connections to a third party
 - Direct observation
- Evaluate the worst-case scenario for attacker
 - Use only PEX and direct observation
 - Vary number of monitoring hosts (rogue clients)
 - Vary peers being monitored (random/most connected)

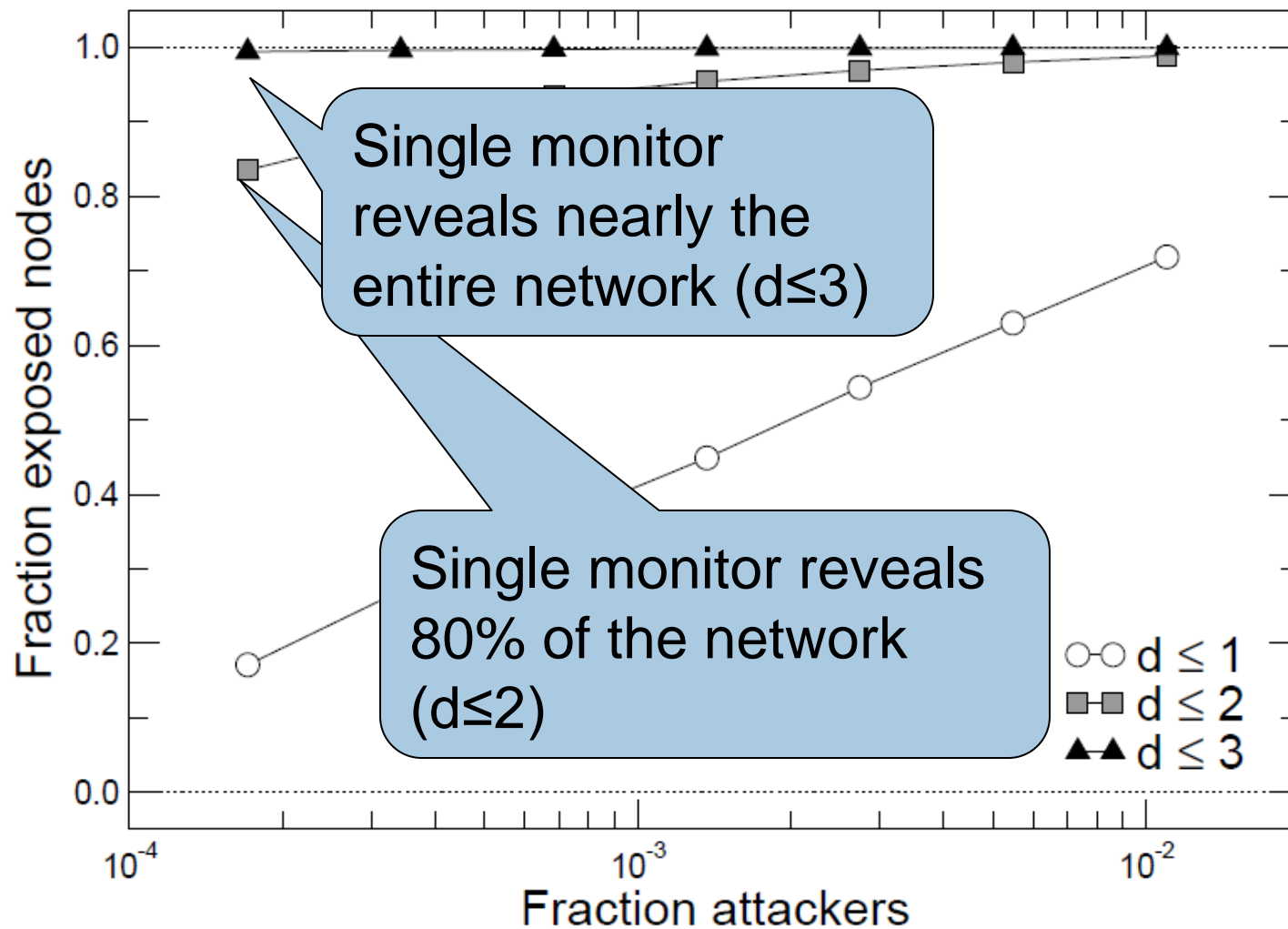
Discovering the Connection Graph

- Randomly select peers to monitor



Discovering the Connection Graph

- Select the **most connected** peers to monitor



Identifying Communities with Partial Graphs

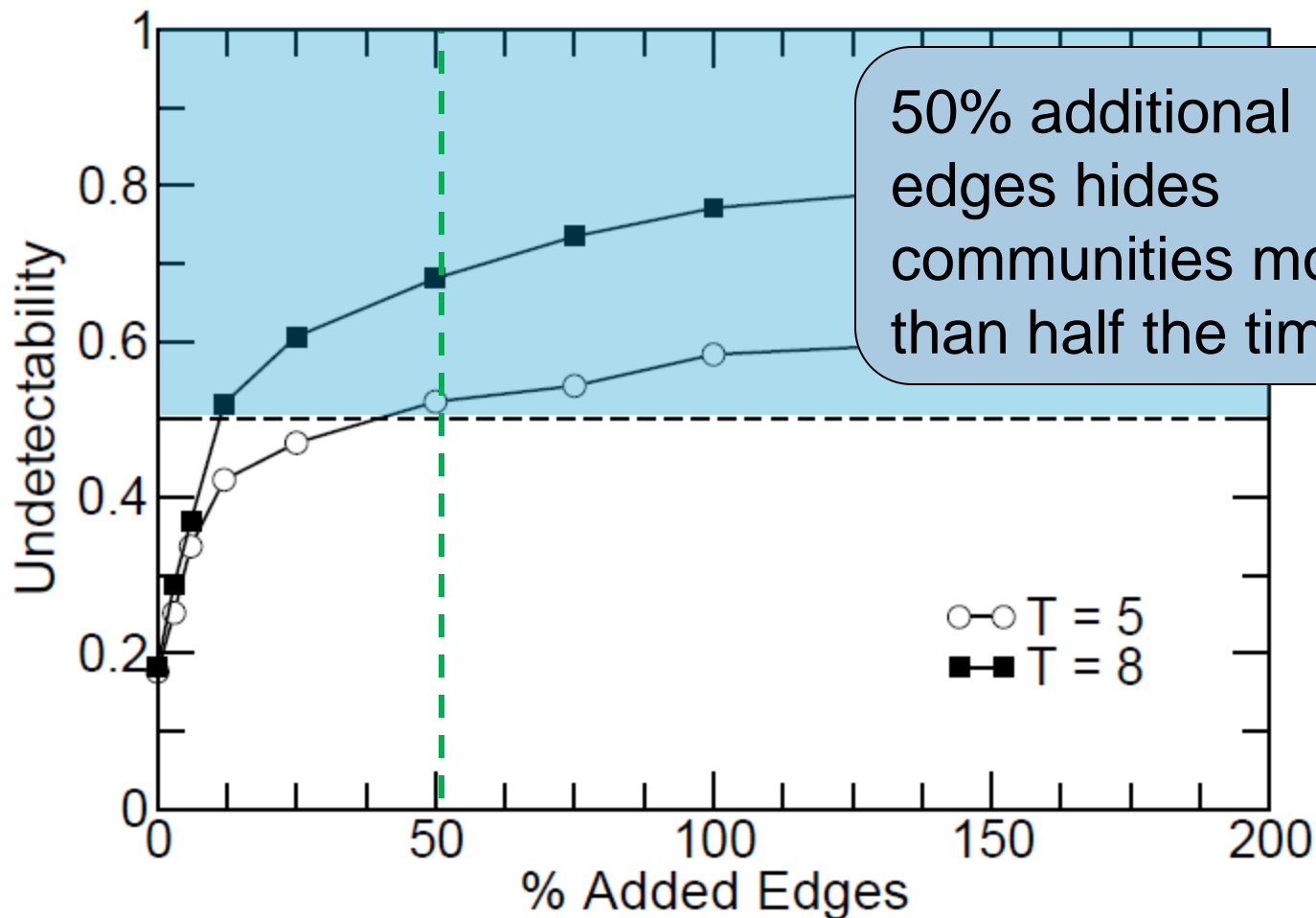
- Monitoring reveals most (but not all) of the network
 - What can be inferred from these partial views?
 - How reliable are these inferences?
- Reliable community inferences
 - Determine probability that node is classified in *partial network* given that it is in the *full network*
 - Run extremal optimization R times
 - How many times (τ) do communities overlap?
- Results (partial)
 - $\tau = 8$, 0.01% monitored, $d \leq 3$: correct 85% of the time
 - $\tau = 8$, 1% monitored, $d \leq 2$: correct 86% of the time

Disrupting Community Identification

- Key assumptions for guilt by association
 - Connections == shared interest
 - Strong communities (relatively low noise in graph)
- To preserve privacy, attack the assumptions
 - Add random connections
 - Number proportional to real ones
- How well does this work?
 - **Undetectability**: How well it hides communities
 - **Deniability**: How many detected communities are wrong

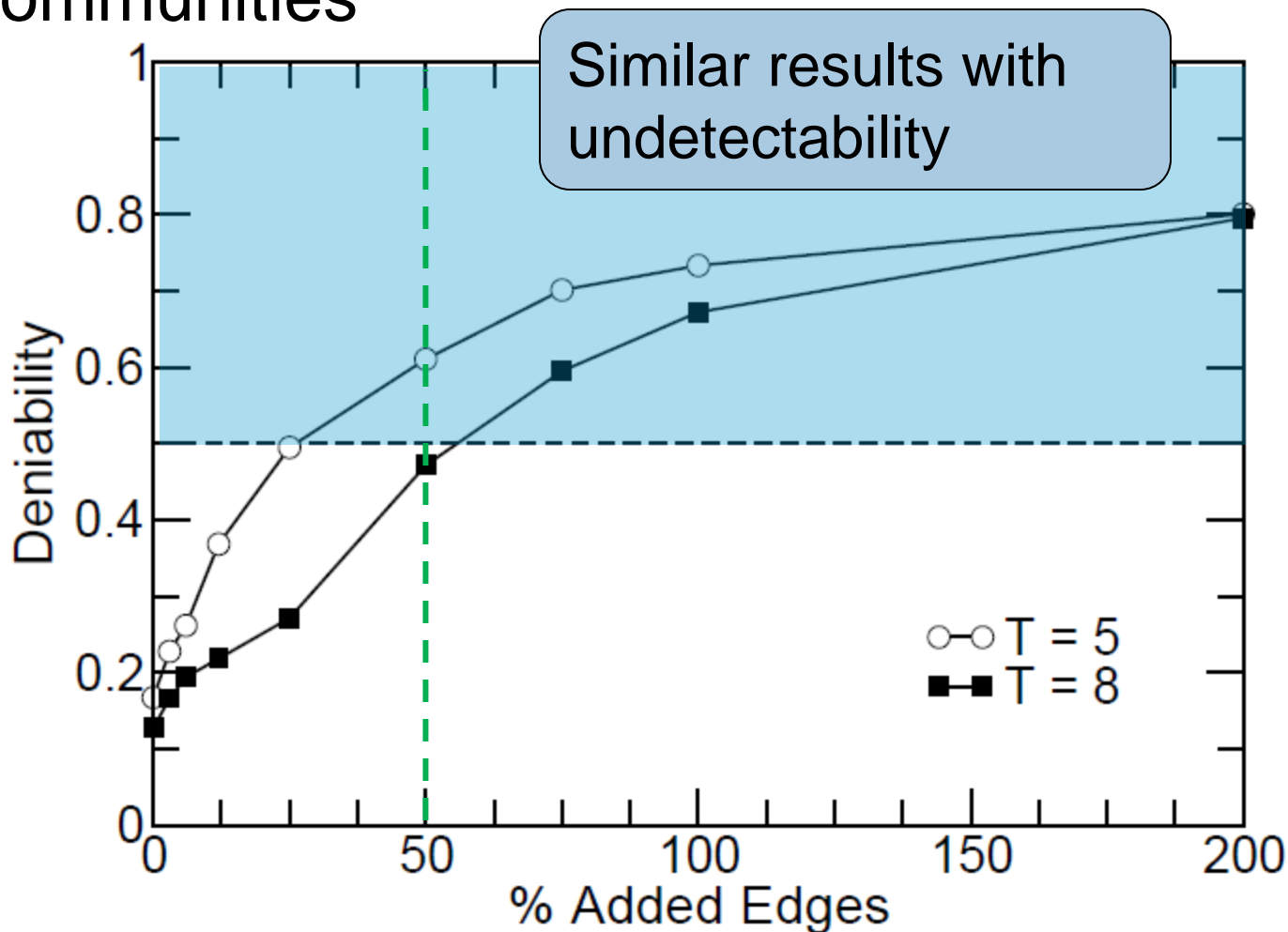
Undetectability

- Percent of time nodes *not classified* into communities



Deniability

- Percent of time nodes *incorrectly* classified into communities



Conclusion

- Communities in BitTorrent
 - Strong communities naturally form
 - Can be exploited using guilt by association
 - Permits lightweight monitoring of BitTorrent
- Disrupting community identification
 - Proposed and evaluated potential solution
 - Adding random edges effectively mitigates threat

- Is this really practical?
 - Where do you get random connections?
 - How much overhead is this?
- SwarmScreen
 - Use real torrents selected at random
 - Cover traffic contributes to real BT swarms
 - Users can control privacy/performance overhead
- Deployed for Vuze BitTorrent client
 - Come see the demo after the talk