Client-side Web Mining for Community Formation in Peer-to-Peer Environments

Kun Liu, Kanishka Bhaduri, Kamalika Das,
Phuong Nguyen and Hillol Kargupta

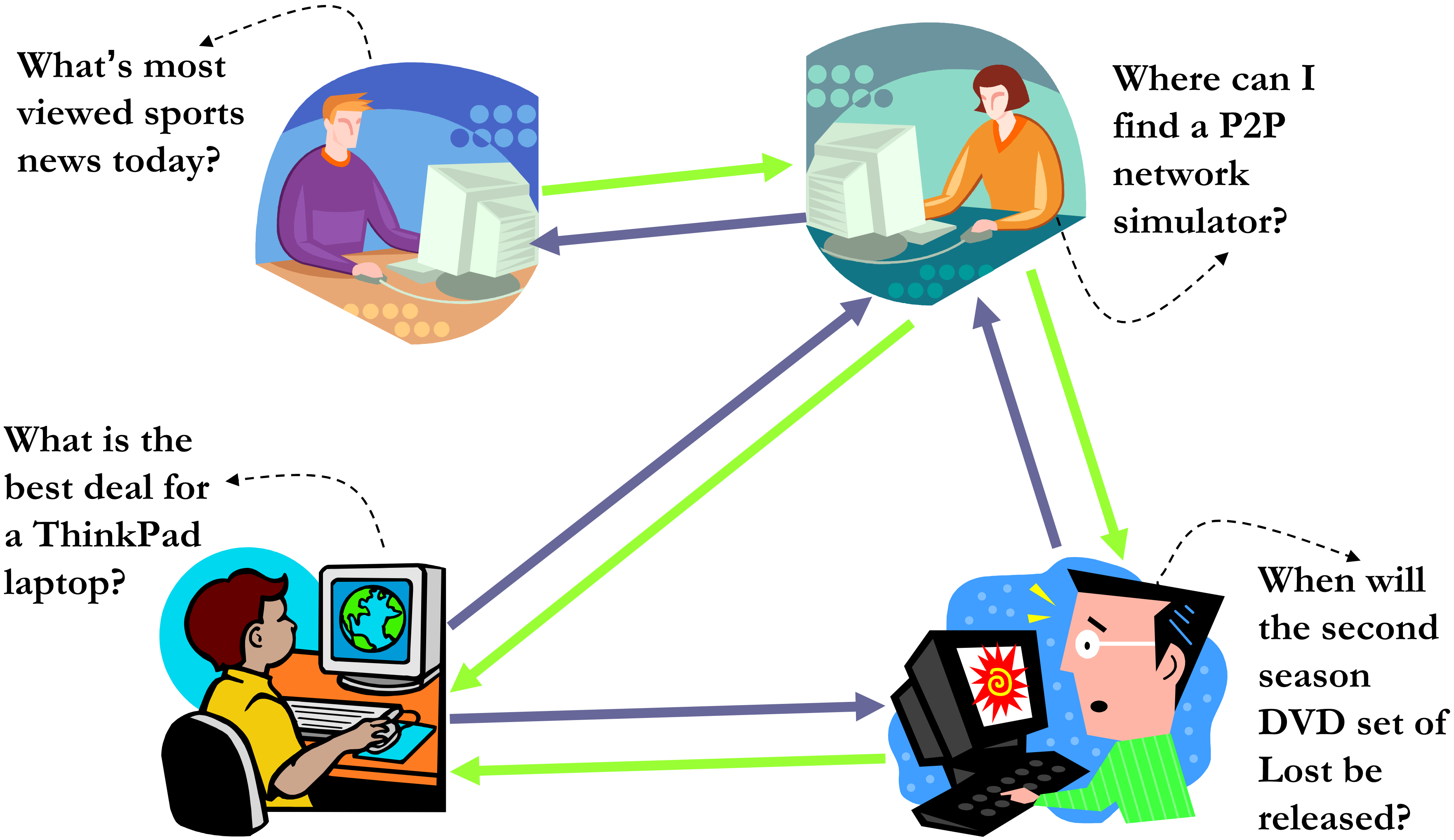
University of Maryland, Baltimore County

WebKDD'06, August 20, 2006, Philadelphia, PA, USA

Motivation

- Online Communities
  - Social motive drives people to seek contact with others
  - Google, Yahoo newsgroups, mailing lists, online forums
  - Most of online communities are under certain central control
- Peer-to-Peer Network
  - SETI, KaZaA, BitTorrent, Gnutella, Napster
- Interest-based Peer-to-Peer Communities
  - A collection of peers in the network that share common interests
  - Self-organizing, no central management
  - Facilitating knowledge sharing
  - Reducing network load

Peer-to-Peer Community



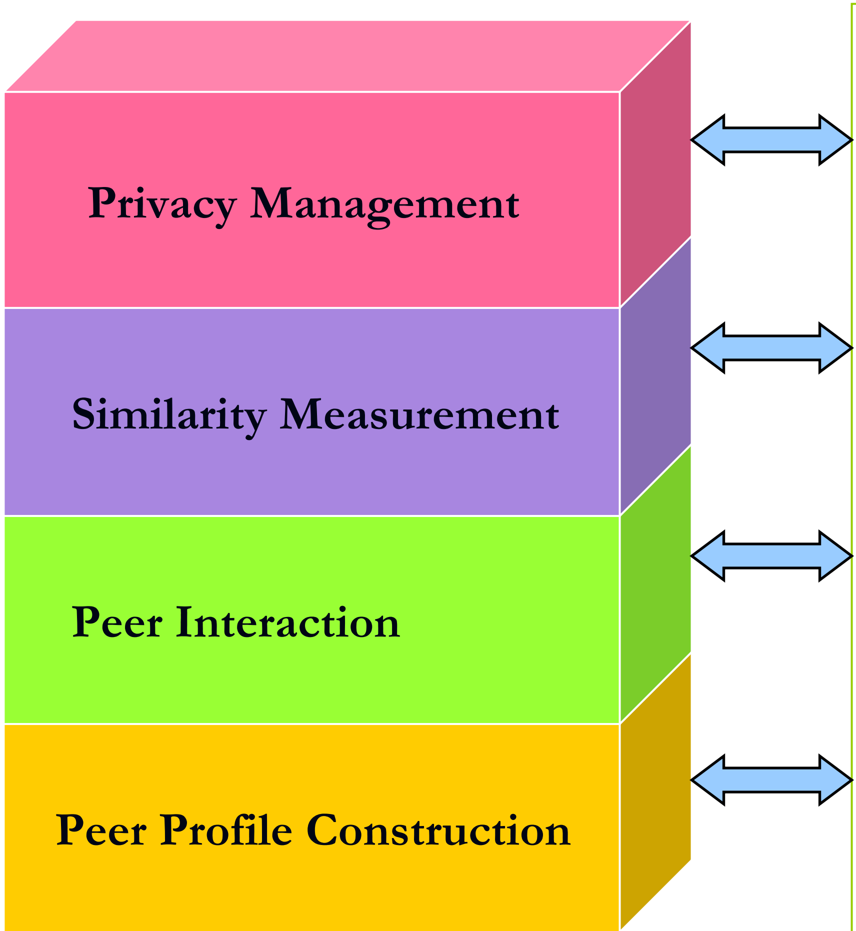
Our Work

- A framework for forming interest-based Peer-to-Peer communities
- Order statistics-based approach to construct communities with hierarchical structures
- Cryptographic protocols to measure similarity between peers without disclosing their personal profiles to each other

Related Work

- Trust-based approach [Wang04]
- Link analysis-based approach [Flake02]
- Ontology matching-based approach [Castano05]
- Attribute similarity-based approach [Khambatti02]

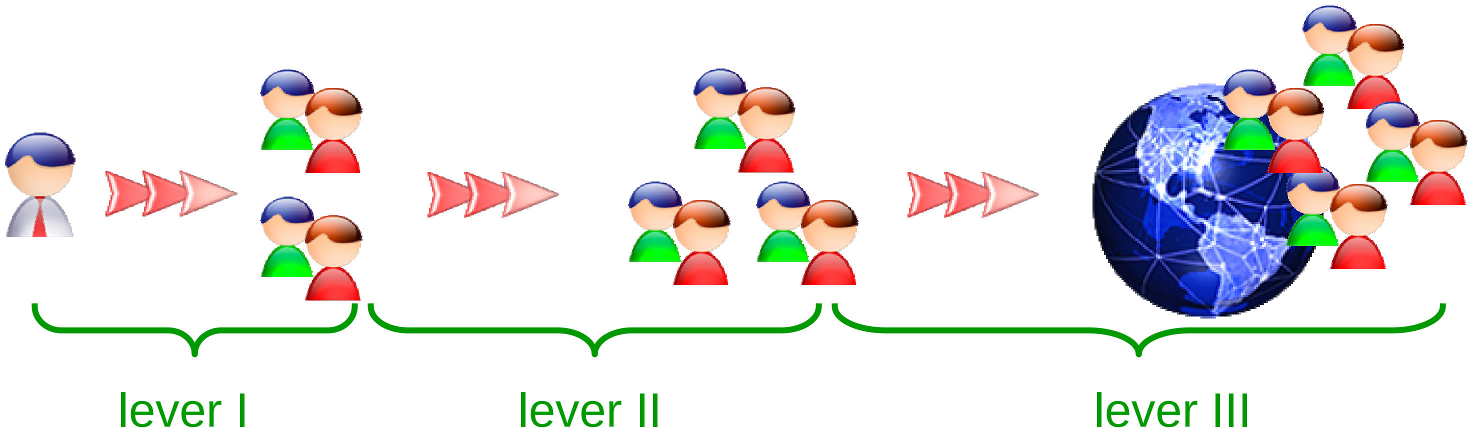
Building Blocks



- Cryptographic protocols are adopted to measure similarity between peers without disclosing their personal profiles
- Inner product between profile vectors used as similarity index. Order statistics-based approach used to build communities with hierarchical structures
- Peer interacts with others by submitting discovery queries to identify potential members; or by replying incoming queries to decide whether it can join a community
- Each peer is associated with a profile vector that represents its interests, e.g., frequencies of web domains a peer has visited

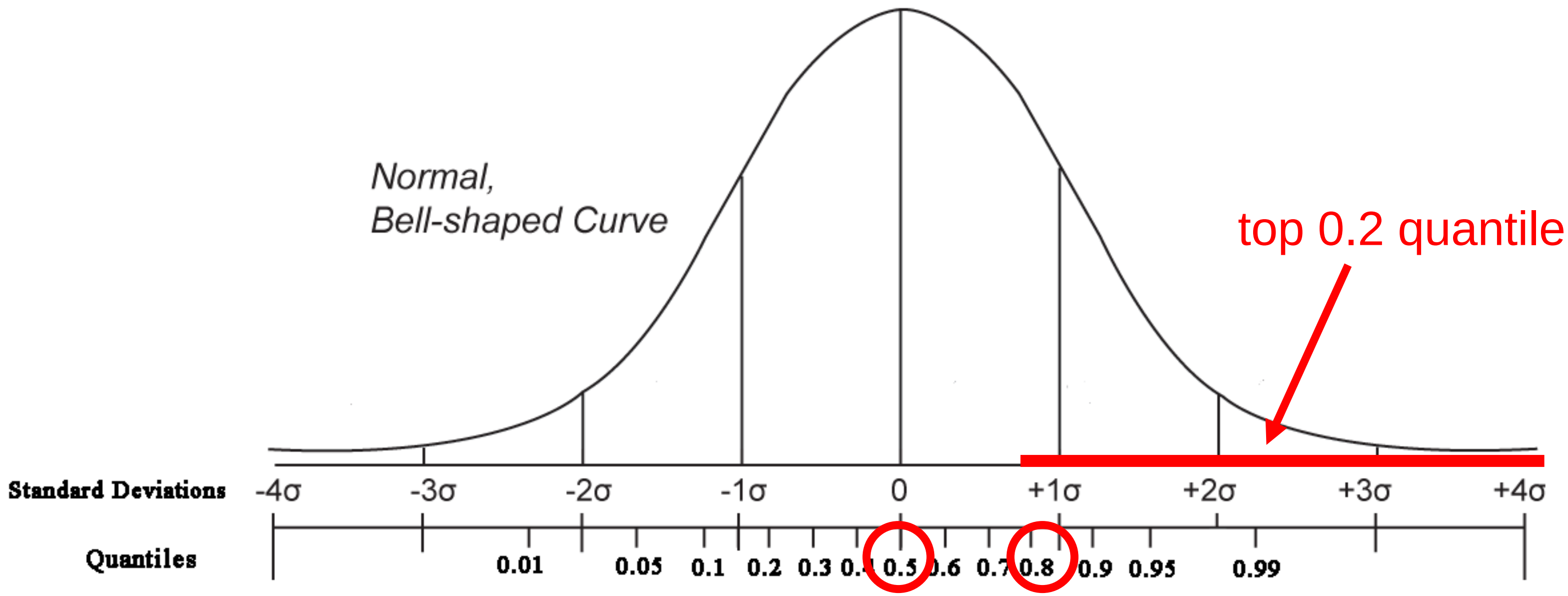
Similarity Measurement

- What is "*similar*" ?
  - We need statistical metric to quantify the similarity
- Hierarchical Structure of the Community



Order Statistics – Distribution-Free Confidence Interval for Quantiles

- Population Quantile
  - Let **X** be a continuous random variable
  - Let ξ_p be the population quantile of order p, *i.e.*, $\Pr\{x \le \xi_p\} = p$



Order Statistics – Distribution-Free Confidence Interval for Quantiles

- Population Quantile Estimation
  - Let **X** be a continuous random variable
  - Let ξ_p be the population quantile of order p, *i.e.,* $\Pr\{x \le \xi_p\} = p$
  - *Let* $x_1 < x_2 < \ldots < x_N$ be N independent samples from **X**
  - We have

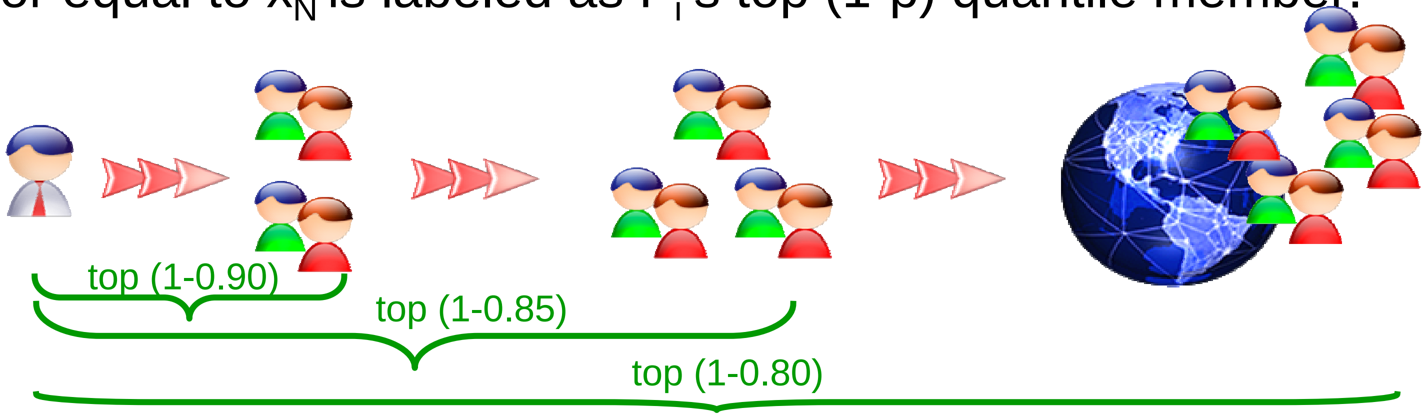
$$\Pr\{x_N > \xi_p\} > q \Rightarrow N \ge \left\lceil \frac{\log(1-q)}{\log p} \right\rceil$$

  - Example:

p (order of quantile)	q (confidence level)	N (sample size)
0.90	0.95	29
0.85	0.95	19
0.80	0.95	14

Quantile Estimation in Network

- The community initiator P_i invokes N random walks (Metropolis-Hastings Sampling) over the network to find N sample peers.
- P_i computes the inner product of his profile vector with each of the sample peers.
- The largest inner product x_N is used as the threshold for estimating quantile ξ_p.
- Any peer in the network whose inner product with P_i is greater than or equal to x_N is labeled as P_i's top (1-p) quantile member.



Privacy Management

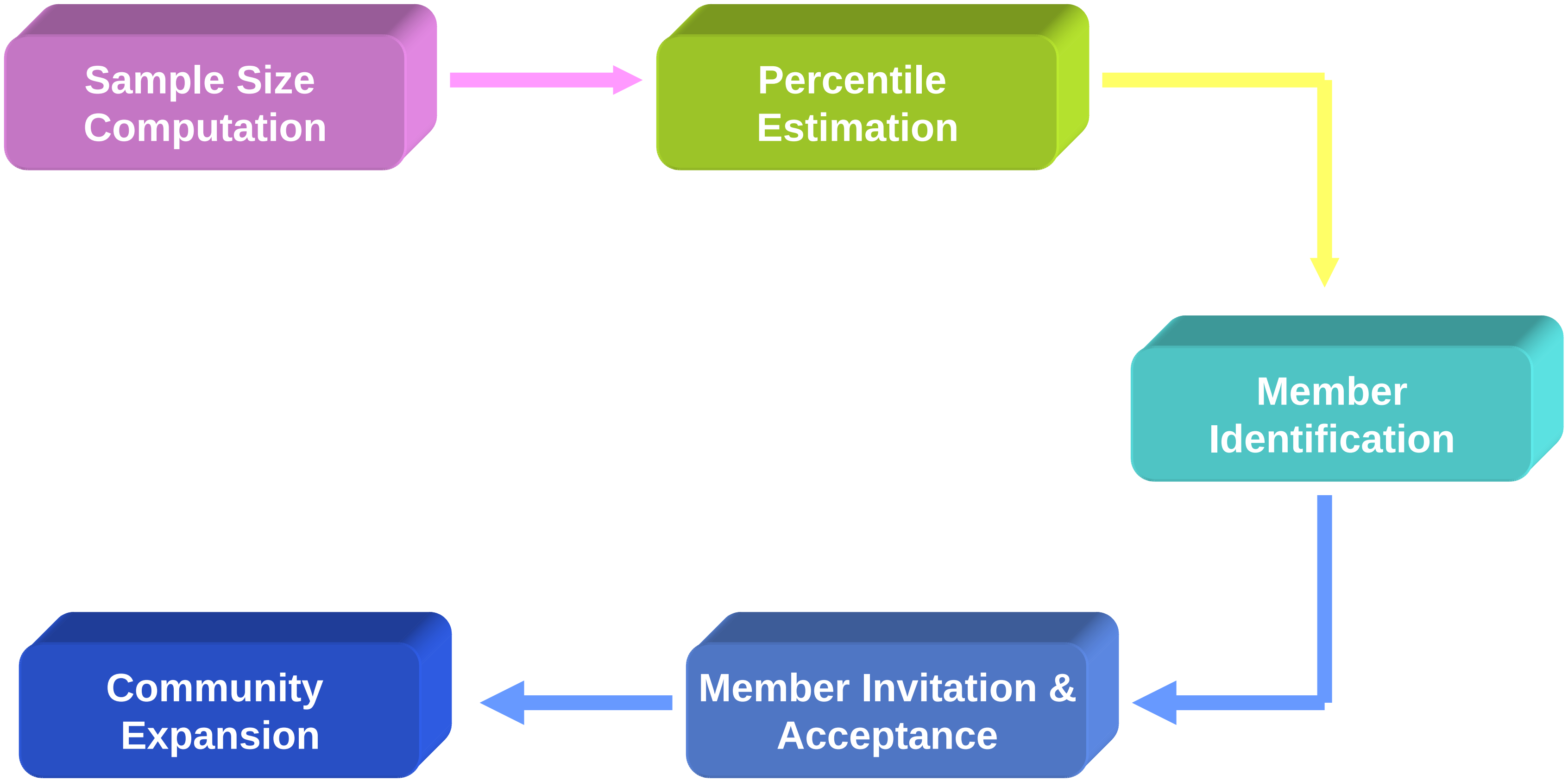
- Private Inner Product Computation
  - To compute the inner product of two profile vectors owned by two different peers, so that neither peer should learn anything beyond what is implied by the peer's own vector and the output of the computation.
- Protocol

Protocol 5.3.1 Private Scalar Product

Private Input of Alice: Vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{Z}_\mu^d$
Private Input of Bob: Vector $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{Z}_\mu^d$
Output of Alice: $\mathbf{x} \cdot \mathbf{y} \bmod \mu$

1: Alice generates a private and public key pair (sk, pk), and sends pk to Bob.
2: For each $i, i = 1, \ldots d$, Alice generates a random number $r_i \in Z_\mu$, and sends $c_i = E_{pk}(x_i, r_i)$ to Bob.
3: Bob computes $w = \prod_{i=1}^{d} c_i^{y_i} \bmod \mu^2$ and sends w back to Alice.
4: Alice computes $\mathbf{x} \cdot \mathbf{y} \bmod \mu = D_{sk}(w)$.

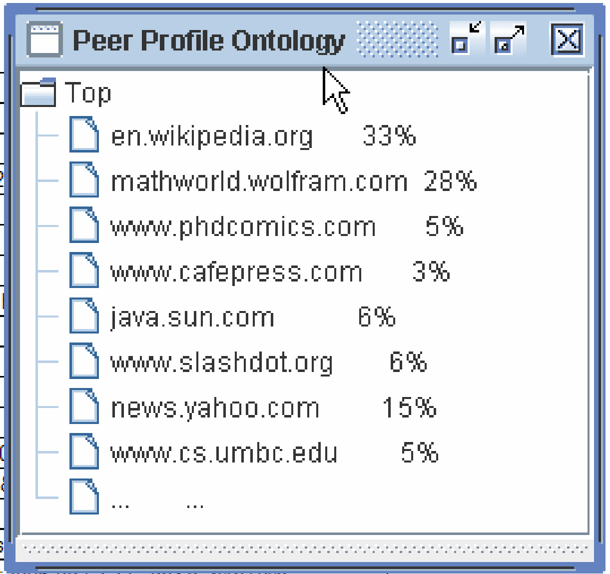
Community Formation Process



Experiments

- Data Collection
  - 15 volunteers from UMBC and JHU
  - 97,050 web browsing history records, 722 unique domains
- Network Topology Generation
  - BRITE: a universal topology generator from Boston University
  - Barabasi model to simulate Internet topology
- Distributed Computation Simulator
  - Distributed Data Mining Toolkit (DDMT) from UMBC

Data Collection



:2006080720060814: kunliu1@http://video.google.com/videoplay?docid=7632211729087286881&hl=en
:2006080720060814: kunliu1@:Host: webmail.umbc.edu
:2006080720060814: kunliu1@:Host: www.umbc.edu
:2006080720060814: kunliu1@http://www.google.com/search?hl=en&lr=&rls=GGLG%2CGGLG%3A2006-09%2
:2006080720060814: kunliu1@http://www.cs.berkeley.edu/~jfc/papers/02/IEEESP02.pdf
:2006080720060814: kunliu1@http://bbs.qq.com/cgi-bin/bbs/show/title?groupid=122:11232&st=&sc
:2006080720060814: kunliu1@http://video.google.com/videoplay?docid=3611345865682027477&hl=en
:2006080720060814: kunliu1@http://news.phoenixtv.com/phoenixtv/72620543991349248/news/20060812/dbgc.l
:2006080720060814: kunliu1@:Host: www.cs.berkeley.edu
:2006080720060814: kunliu1@:Host: www.informatik.uni-trier.de
:2006080720060814: kunliu1@:Host: kdd.ics.uci.edu
:2006080720060814: kunliu1@:Host: www.vacancies.auckland.ac.nz
:2006080720060814: kunliu1@http://by105fd.bay105.hotmail.msn.com/cgi-bin/HoTMaiL?fti=yes&curmbox=000
:2006080720060814: kunliu1@https://webmail.umbc.edu/src/download.php?absolute_dl=true&passed_id=7046
:2006080720060814: kunliu1@http://www.ics.uci.edu/~pazzani
:2006080720060814: kunliu1@ftp://ftp.ics.uci.edu/pub/machine-learning-databases/undocumented/connectionis
:2006080720060814: kunliu1@http://www.google.com/search?sourceid=navclient&ie=UTF-8&rls=GGLG,GGLG:
:2006080720060814: kunliu1@http://zmen001.spaces.live.com/?
:2006080720060814: kunliu1@http://royliuk.spaces.live.com/?
:2006080720060814: kunliu1@:Host: www.sunrain.net
:2006080720060814: kunliu1@ftp://ftp.ics.uci.edu/pub/machine-learning-databases/image/segmentation.data
:2006080720060814: kunliu1@http://www.cs.auckland.ac.nz/phd/grads.html
:2006080720060814: kunliu1@:Host: www.cs.auckland.ac.nz
:2006080720060814: kunliu1@http://www.google.com/search?q=random+projection+privacy&hl=en&lr=&rls=GGLG,GGLG:2006-
:2006080720060814: kunliu1@http://www.cs.unc.edu/~lin
:2006080720060814: kunliu1@file:///C:/MATLAB701/work/experiments/realdata/backup/pcaAttack.m
:2006080720060814: kunliu1@http://by105fd.bay105.hotmail.msn.com/cgi-bin/HoTMaiL?fti=yes&curmbox=00000000%2d0000%2d0000%2d0000%
:2006080720060814: kunliu1@http://www.umbc.edu/orientation/freshmen/chat.html
:2006080720060814: kunliu1@http://www.cs.umbc.edu/~kunliu1/music/onlytime.wma
:2006080720060814: kunliu1@http://www.ics.uci.edu/~welling
:2006080720060814: kunliu1@http://hkn.berkeley.edu/student/CourseSurvey/instructors/CS/TA/Duan,Yitao
:2006080720060814: kunliu1@http://www.google.com/search?q=random+projection&hl=en&lr=&rls=GGLG,GGLG:2006-09,GGLG:en&start=20&sa=N
:2006080720060814: kunliu1@ftp://ftp.ics.uci.edu/pub/machine-learning-databases/SUMMARY-TABLE
:2006080720060814: kunliu1@http://www.vacancies.auckland.ac.nz/positiondetail.asp?p=4383

Network Topology

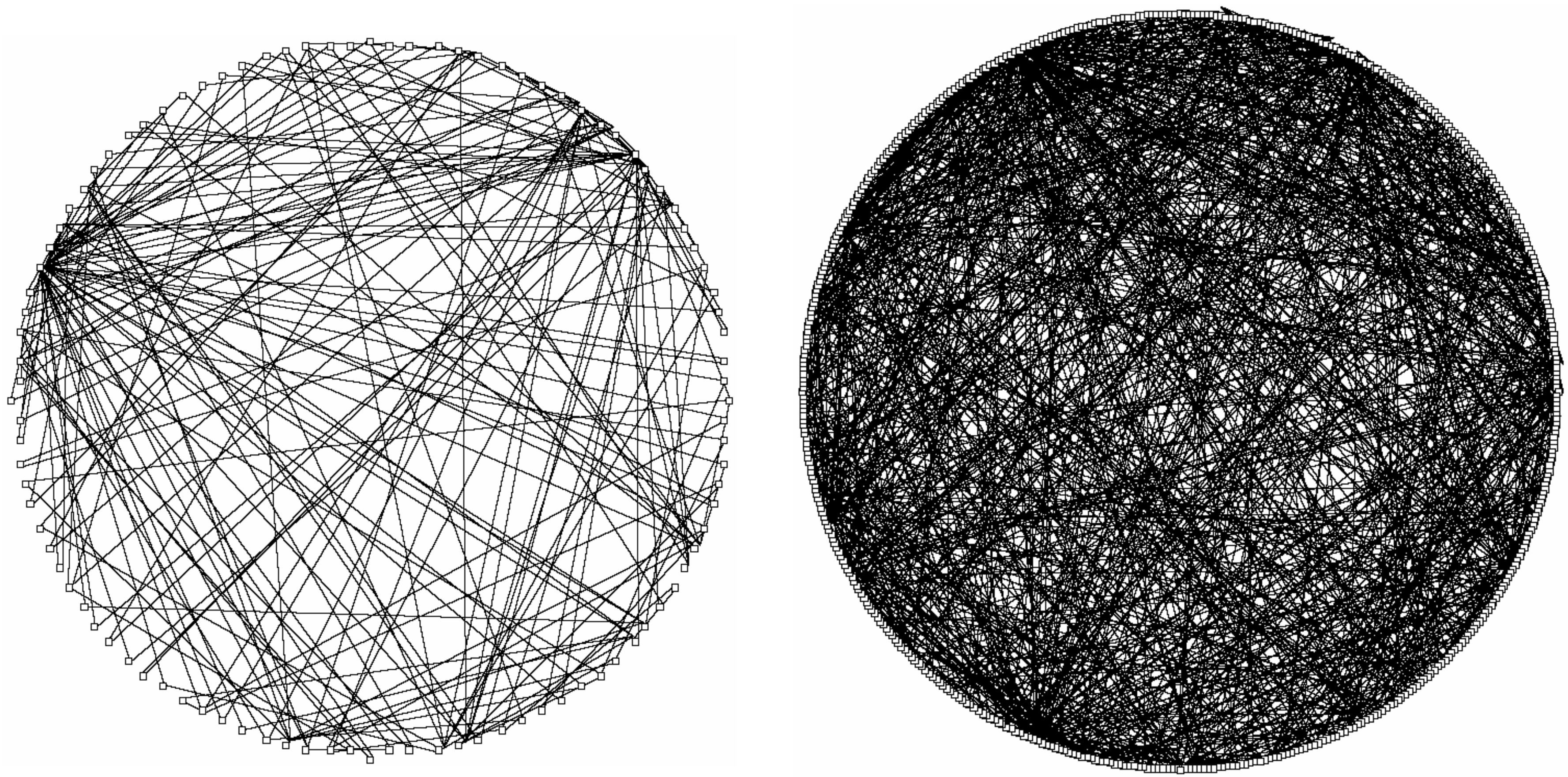
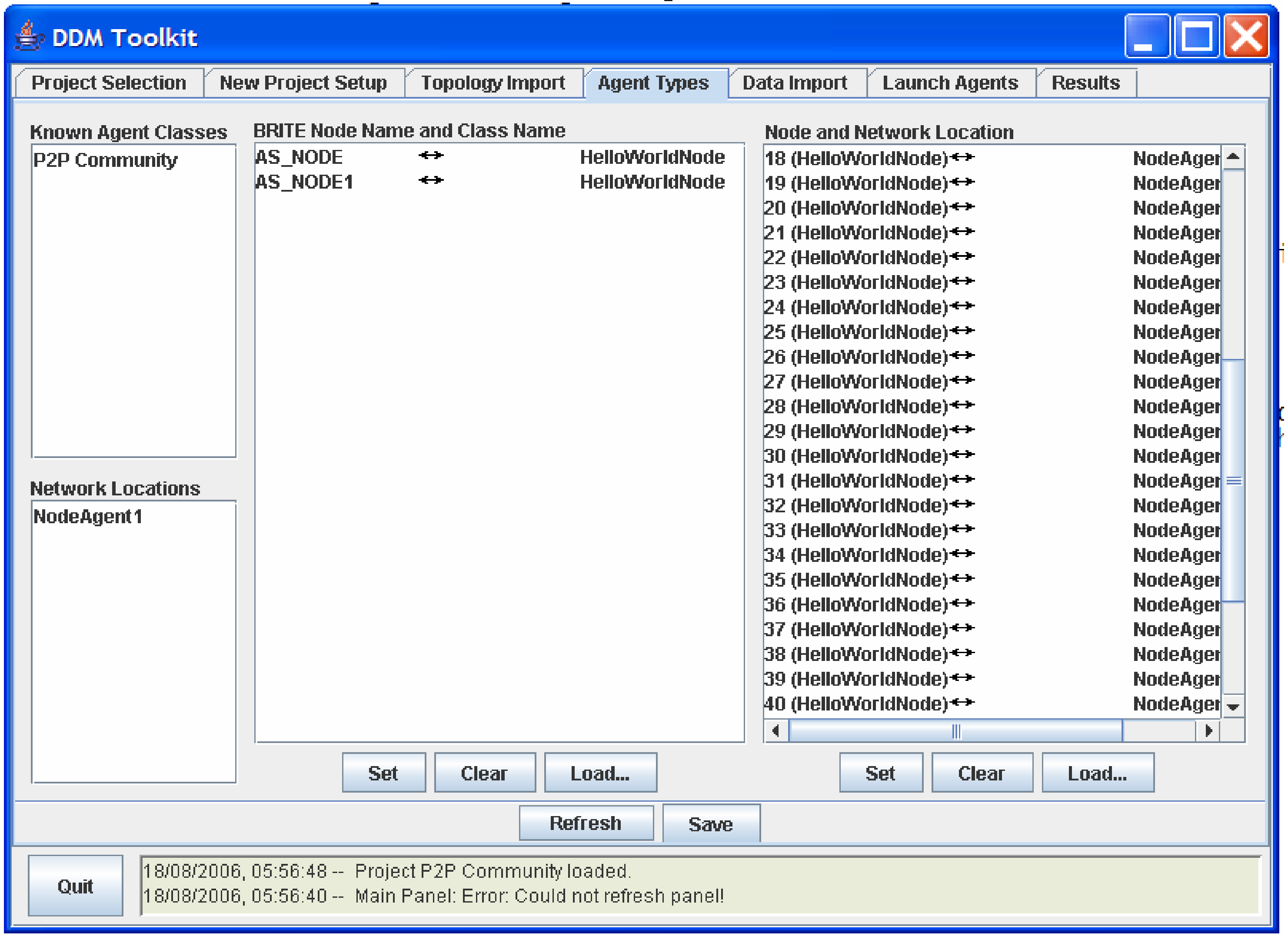


Fig. Topology generated by Barabasi model with BRITE. Left: 100 nodes; Right: 500 nodes.

Distributed Computation Simulator



Experiments of Population Quantile Estimation

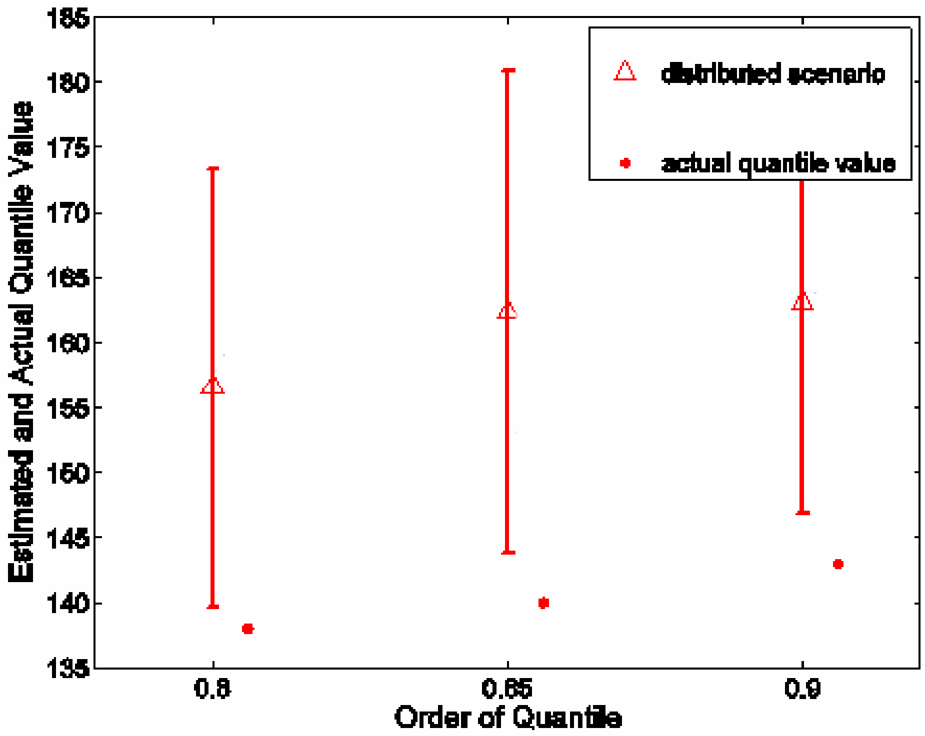


Fig.1: Estimated and actual quantile value w.r.t. the order of quantiles. The results are an average of 100 independent runs.

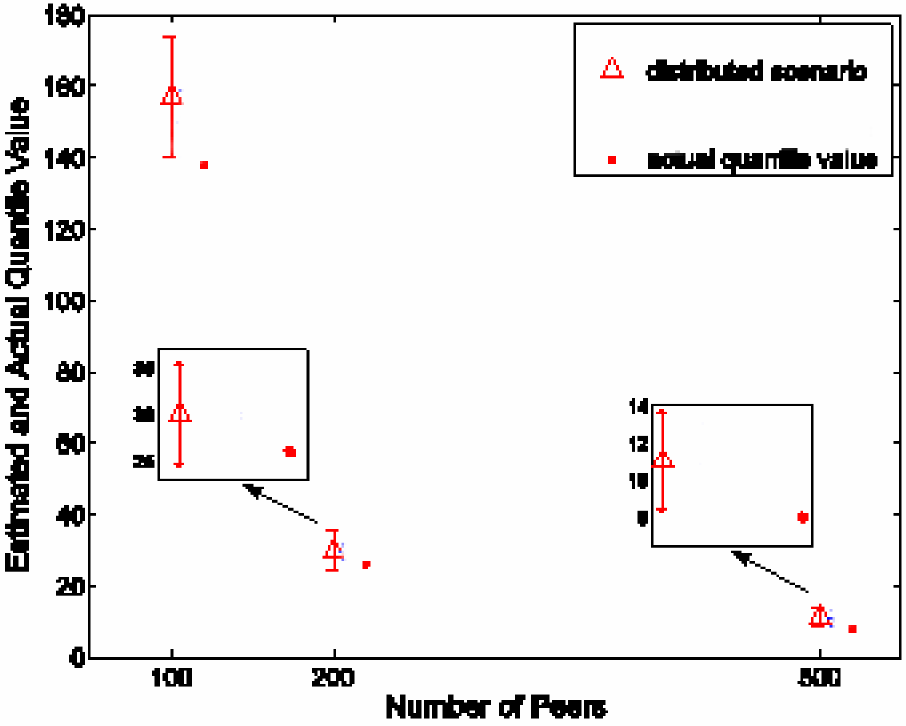


Fig. 2: Estimated and actual quantile value w.r.t. the number of peers for fixed p=0.8, q=0.95. The results are an average of 100 independent runs.

Experiments of Community Formation

TTL	Ave Num of Community Members	Time (in secs)
3	3	55.00
4	8	77.50
8	13	173.00

Fig. 4: Average number of community members found by a peer without community expansion. 95% confidence, 80% quantile, 100 peers in total.

TTL	Ave Num of Community Members	Time (in secs)
3	7	59.00
4	12	82.50
8	17	179.00

Fig. 5: Average number of community members found by a peer with community expansion. 95% confidence, 80% quantile, 100 peers in total.

Future Work

- New approach to build peer's profile
- Experiments in a real distributed environment

References

- [Castano05] S. Castano and S. Montanelli. Semantic self-formation of communities of peers. In Proceedings of the ESWC Workshop on Ontologies in Peer-to-Peer Communities, Heraklion, Greece, May 2005.
- [Khambatti02] M. Khambatti, K. D. Ryu, and P. Dasgupta. Efficient discovery of implicitly formed peer-to-peer communities. International Journal of Parallel and Distributed Systems and Networks, 5(4):155–164, 2002.
- [Wang04] Y. Wang and J. Vassileva. Trust-based community formation in peer-to-peer file sharing networks. In Proceedings IEEE International Conference on Web Intelligence (WI'04), pages 341–338, Beijing, China, October 2004.
- [Flake02] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self organization and identification of web communities. IEEE Computer, 35(3):66–71, March 2002.
- [BRITE] http://www.cs.bu.edu/brite/
- [DDMT] http://www.umbc.edu/ddm/wiki/software

Thank You!
Questions?