Small-World File-Sharing Communities

Presented by: Xiaolong Jin

Based on the paper: Adriana Iamnitchi, Matei Ripeanu, and Ian Foster, Small-World File-Sharing Communities, Infocom 2004, Hong Kong, March 2004,
available at: http://people.cs.uchicago.edu/~anda/

Introduction

- Observation: Large-scale, Internet-based distributed system are hard to manage.
  - For example: For a resource-sharing system, its challenges lie in:
    - Ad-hoc network
    - Intermittent resource participation
    - Large and variable scale
    - High failure rate, etc.
- Problem: How to optimize such systems?
- Solution: To (1) **understand their user behavior**, and then (2) design efficient mechanisms.

Intuition

- Observations from real networks:
  - The popularity of Web pages follows a **Zipf** distribution
  - Node degrees of many networks are distributed according to a **power law**
  - Many networks form **small-world** topologies
- Intuitive questions:
  - **Q1**: Are there any patterns in the way scientists share resources that could be exploited for designing mechanisms?
  - **Q2**: Are these patterns typical of scientific communities or are they more general?

The Data-Sharing Graph

- The data-sharing graph captures the virtual relationship among users who request the same data at around the same time. Specifically,
- Definition: *The data-sharing graph is a graph where nodes are users and an edge connects two users who have similar interests in data.*
- Criterion for similarity: the number of shared requests between to users within a specified time interval

Three Data-Sharing Communities

- The D0 experiment: a high-energy physics collaboration
- The Web observed from the Boeing traces
- The Kazaa peer-to-peer file-sharing system

The D0 Experiment

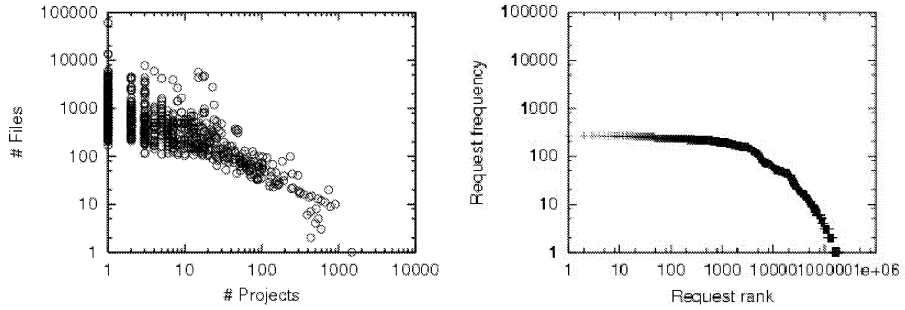


Fig. 1. *Left:* Number of file requests per project in D0. *Right:* File popularity distribution in D0

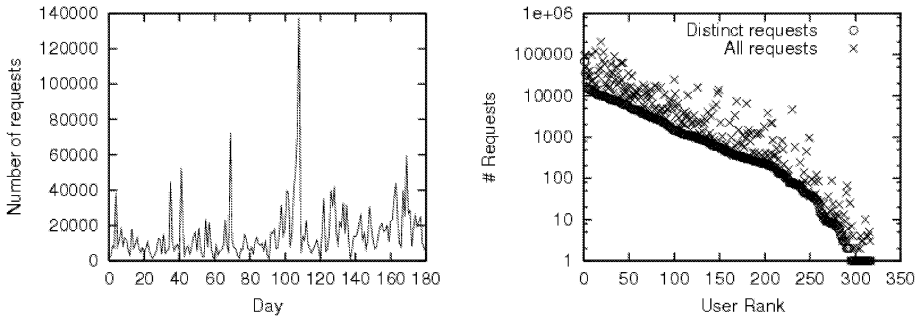


Fig. 2. *Left:* Number of file requests per day in D0. *Right:* Number of files (total and distinct) asked by each user during the 6-month interval.

The Web

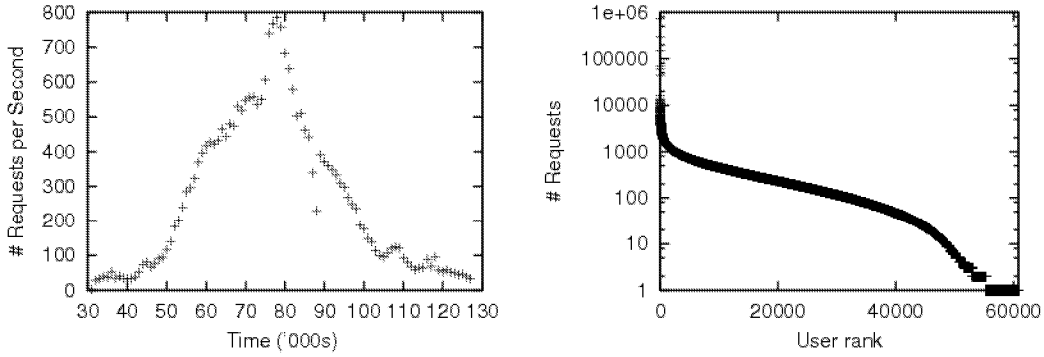


Fig. 3. *Left:* Activity level (averaged over 15-minute intervals). *Right:* Number of requests per Web user.

The Kazaa System

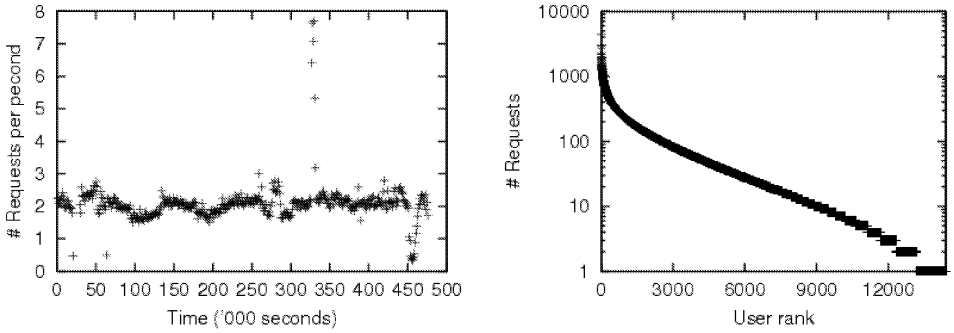


Fig. 4. *Left:* Activity level (averaged over 100 s) in Kazaa; *Right:* Number of requests per user in KaZaa

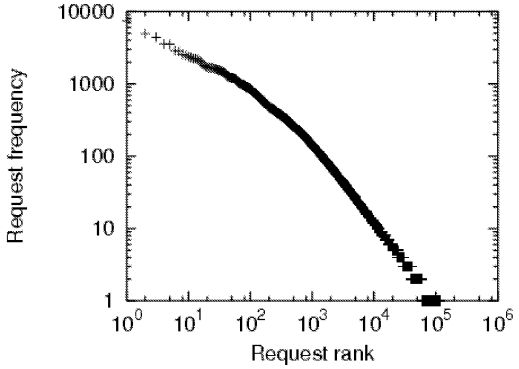
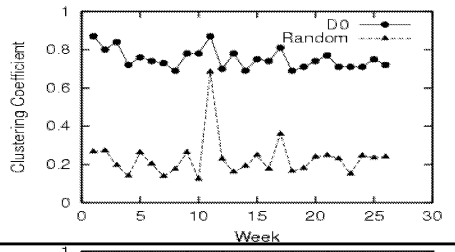


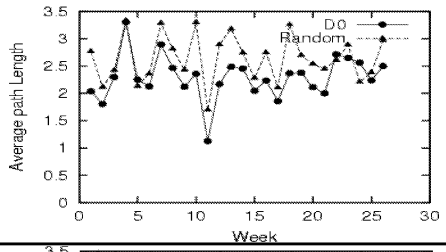
Fig. 5. The file popularity distributions in Kazaa follows Zipf's law.

Small-World Data-Sharing Graphs

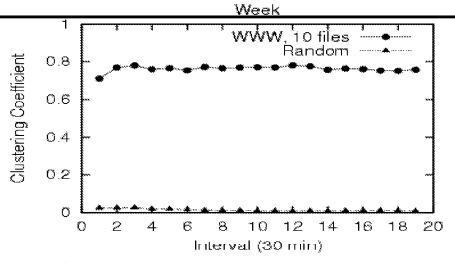
- As compared to a random graph, a small-world graph has:
  - Larger clustering coefficient
  - Smaller average path length
  - *Loosely connected collections of highly connected sub-graphs*

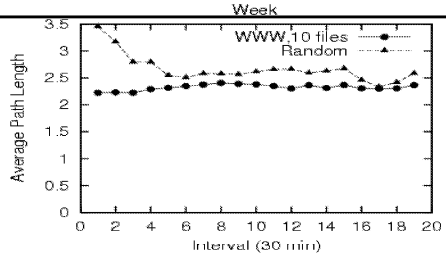
D 0

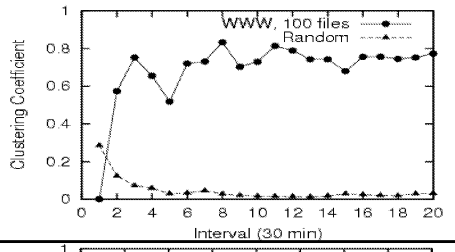


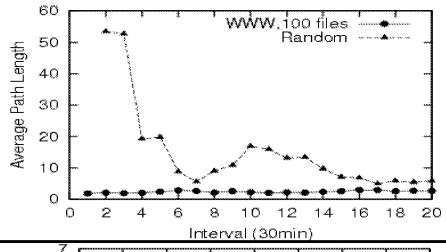


The Web

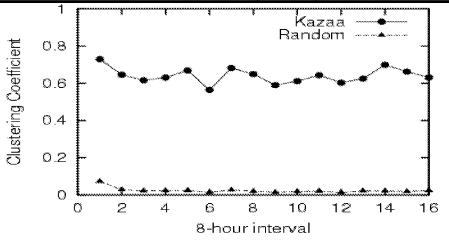


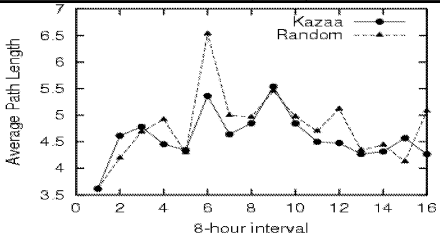


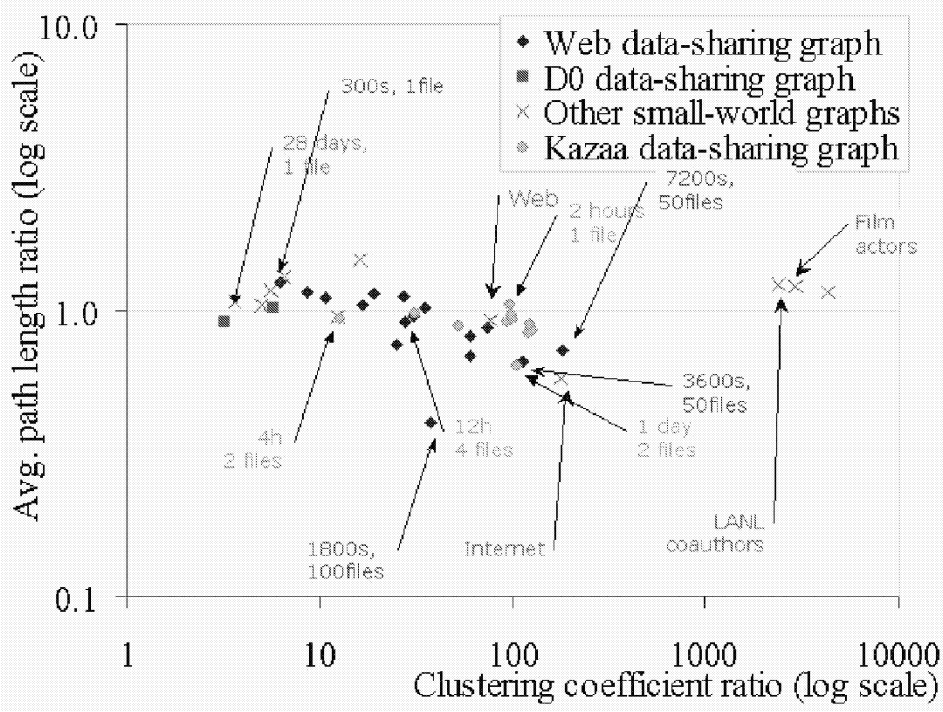




Kazza







- General observations:
  - Data-sharing graphs with different durations and similarity criteria are small-worlds
  - Well connected clusters exists in the communities concerned
  - There is a small average path length between any two nodes in a data-sharing graph

Human Nature or Zipf Law

- **Q3**: Are the small-world properties consequences of previously documented characteristics or do they reflect a new observation concerning users' preferences in data?
  - To examine whether the large clustering coefficient is a natural consequence of the data-sharing graph definition
  - To analyze the influence of time and space locality in file access

Affiliation Networks

- Definition: *an affiliation network is a social network where participants (nodes) in the same interest groups (e.g., clubs, the authors of a paper) are connected.*

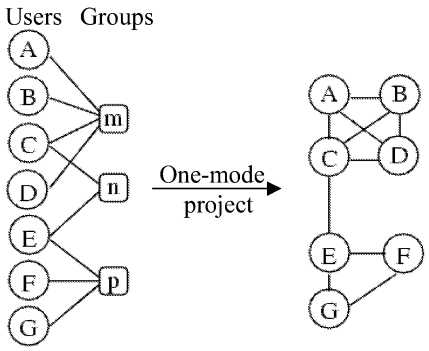


Fig. 15. A bipartite network (left) and its unipartite projection (right). Users A-G access files m-p. In the unipartite projection, two users are connected if they requested the same file.

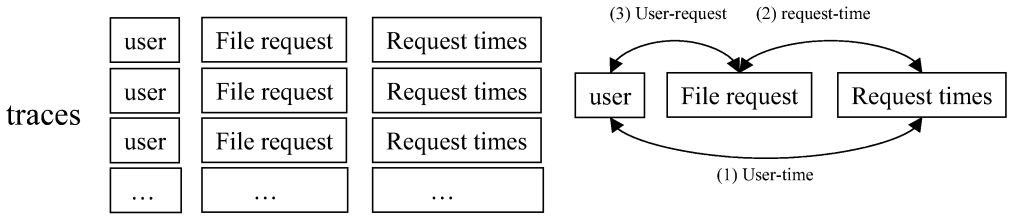
- Comparison: properties of data-sharing graphs, measured and modeled as unimodel projection of affiliation networks

	Interval	Users	Files	Clustering		Average degree	
				Theory	Measured	Theory	Measured
D0	7 days	74	28638	0.0006	0.65	1242.5	3.3
	28 days	151	67742	0.0004	0.64	7589.6	6.0
Web	2 min	3385	39423	0.046	0.63	50.0	22.9
	30 min	6757	240927	0.016		1453.1	304.1
Kazaa	1 h	1629	3393	0.55	0.60	2.9	2.4
	8 h	2497	9224	0.30	0.48	9.5	8.7

- Observations: The large clustering coefficient is not caused by the definition of the data-sharing graph as an one-mode projection of an affiliation network with non-Poisson degree distribution

Zipf Law and Time Locality

- Large clustering coefficient vs. Zipf law and time and space locality
  - Time locality: an item is more popular during a limited interval
- **Q4**: *Are the properties we identified in the data-sharing graph, especially the large clustering coefficient, an inherent consequence of these well-known behaviors ?*
- Means: generate random trances preserving the documented characteristics but break the **user-request** association:



- Three experiments:
  - ST1: Break relationships (1), (2), and (3)
  - ST2: Break relationships (1) and (3) //maintain request-time relationship
  - ST3: Break relationships (1) and (2) // maintain user-time relationship

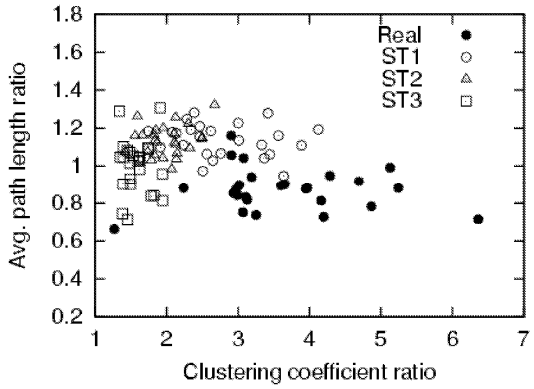


Fig. 20. Comparison of the small-world data-sharing graphs as resulted from the real and synthetic D0 traces.

- Observation: The synthetic data-sharing graphs are still small-worlds (although they are less "small-worldy".)