Analyzing Peer-To-Peer Traffic Across Large Networks

From：IEEE/ACM TRANSACTIONS ON NETWORKING,
VOL. 12, NO. 2, APRIL 2004

Presented by 江文德

Outline

- Introduction
- Methodology
- Characterization Metrics
- Overview of P2P Traffic and System Dynamics
- Traffic Characterization
- Conclusions and Future Work

Introduction

- Two Categories of P2P Traffic
  - signaling
  - data transfer
- Previous Projects Almost Focused on P2P Signaling Traffic
  - They gather traffic by setting up P2P crawler.

Introduction（Cont.）

- Research Questions for P2P System Design and Traffic Engineering
  - How is the P2P traffic distributed across the Internet?
  - What are the characteristics of the application-level P2P network connectivity?
  - How dynamic are the P2P system, both temporally and spatially?

Methodology

- Popular P2P Applications
  - FastTrack, Gnutella, and DirectConnect
- Measurement Approach
  - Offline analysis of flow-level data gathered from multiple routers across a large ISP's backbone
  - IP address, network prefix, autonomous system (AS)
  - Cisco's *NetFlow* service

Methodology（Cont.）

- Advantages
  - It doesn’t require knowledge about the P2P protocol.
  - It approach is non-intrusive and all the traffic data can be collected.
  - It gathers information on both the signaling traffic and the data download traffic.
  - It is conducive to determining the impact of P2P traffic on certain regions of the network.

Methodology（Cont.）

- Limitations
  - We are not able to obtain application-level details.
  - We may not capture the complete flow of traffic.

Characterization Metrics

- Host Distribution
- Traffic Volume
- Host Connectivity

Characterization Metrics (Cont.)

- Traffic Pattern Over Time

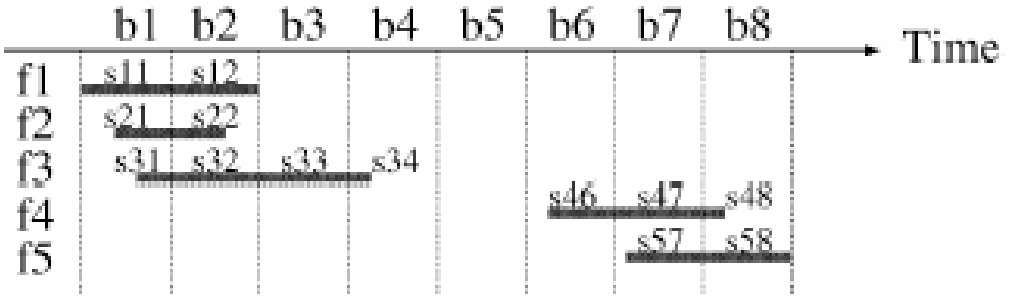


Fig. 1. The binning of *netflow* records.

$$\text{Volume}(b_l) = \sum_{i=1}^{n} \text{Volume}(s_{il}).$$

Characterization Metrics (Cont.)

- Connection Duration and On-Time
  - f_i and f_j are concurrent *iff*

    StartTime (f_j) <= FinishTime (f_i) + δ

    where $i < j$, f_i and f_j are two flows
- Mean Bandwidth Usage

$$\text{Bandwidth}_R(h) = \frac{\text{Volume}_R(h)}{\text{OnTime}_R(h)}$$

where h is a given host

Overview of P2P Traffic and System Dynamics

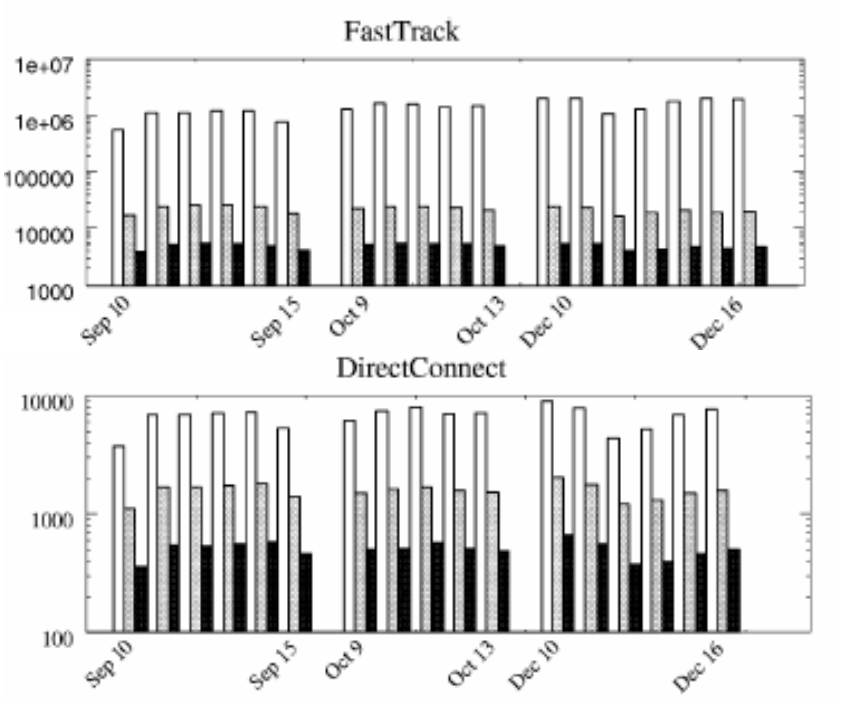
- Summary Statistics for the P2P Data Set

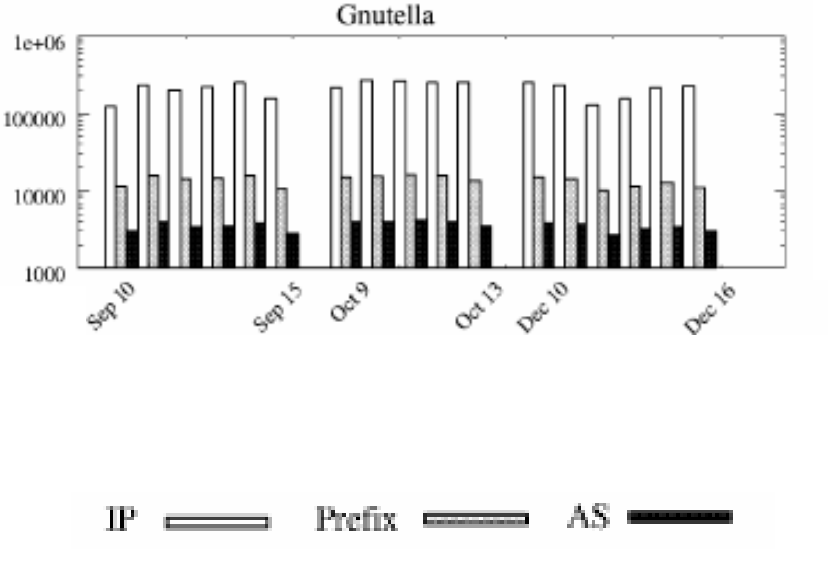
TABLE I
Netflow DATA SET OF P2P TRAFFIC OVER TCP

Date	Protocol	Number of records	Total number of unique IPs	Number of unique IPs per day	Total traffic volume (GBytes/day)	Traffic volume per IP (MBytes/day)
9/10/2001 - 9/15/2001	Gnutella	37,853,281	718,464	197,445	211	2.2
	FastTrack	110,533,024	3,403,900	998,669	773	1.6
	DirectConnect	595,606	22,852	6,244	48	15.4
10/9/2001 - 10/13/2001	Gnutella	49,649,348	823,532	247,114	272	2.2
	FastTrack	184,113,038	4,450,149	1,485,370	1,153	1.6
	DirectConnect	566,740	23,211	7,193	56	15.6
12/10/2001 - 12/16/2001	Gnutella	69,578,723	887,520	236,954	242	2.0
	FastTrack	340,690,074	5,924,072	1,934,460	1,776	1.8
	DirectConnect	701,712	29,925	7,213	71	19.6

Overview of P2P Traffic and System Dynamics（Cont.）

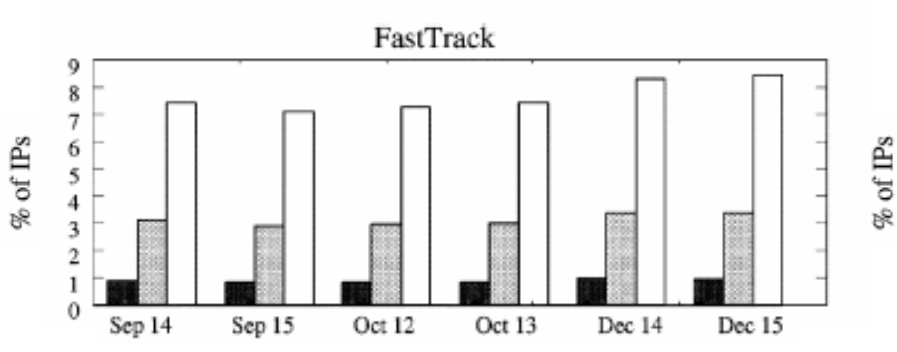
- Host Distribution

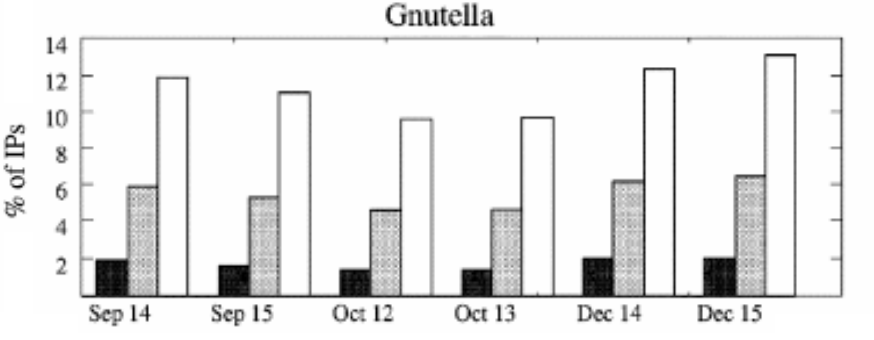


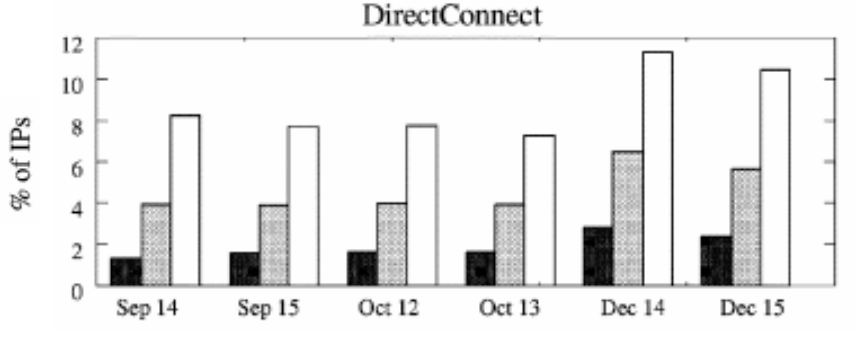


Overview of P2P Traffic and System Dynamics（Cont.）

- Traffic Volume Distribution

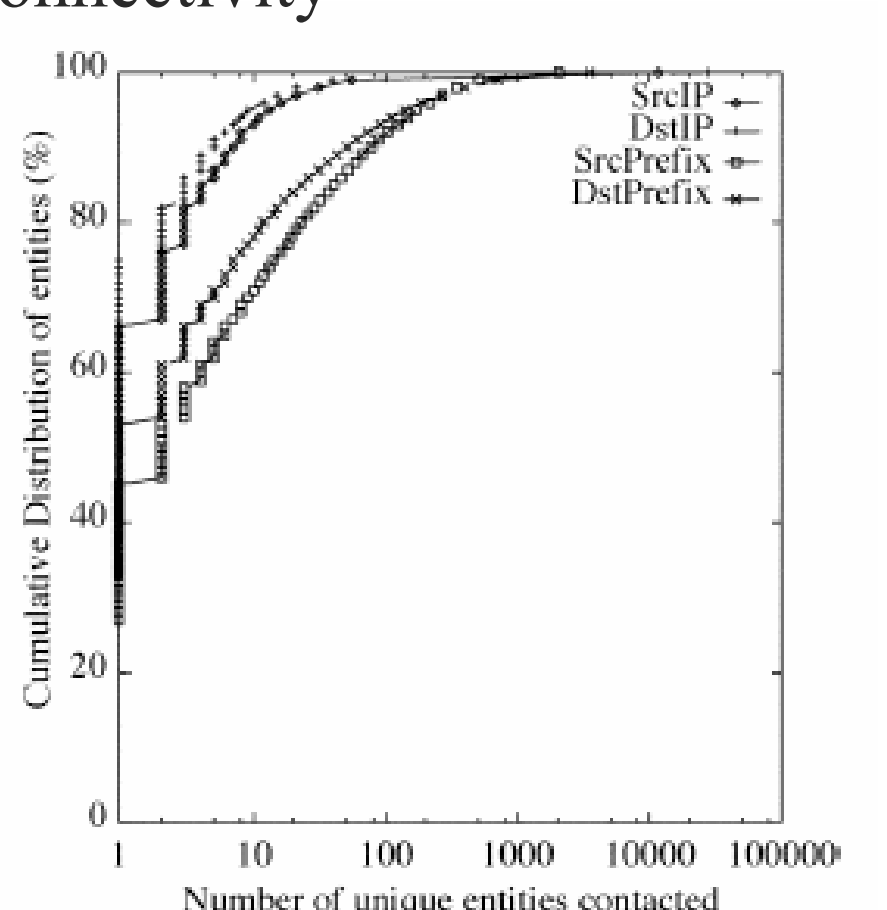






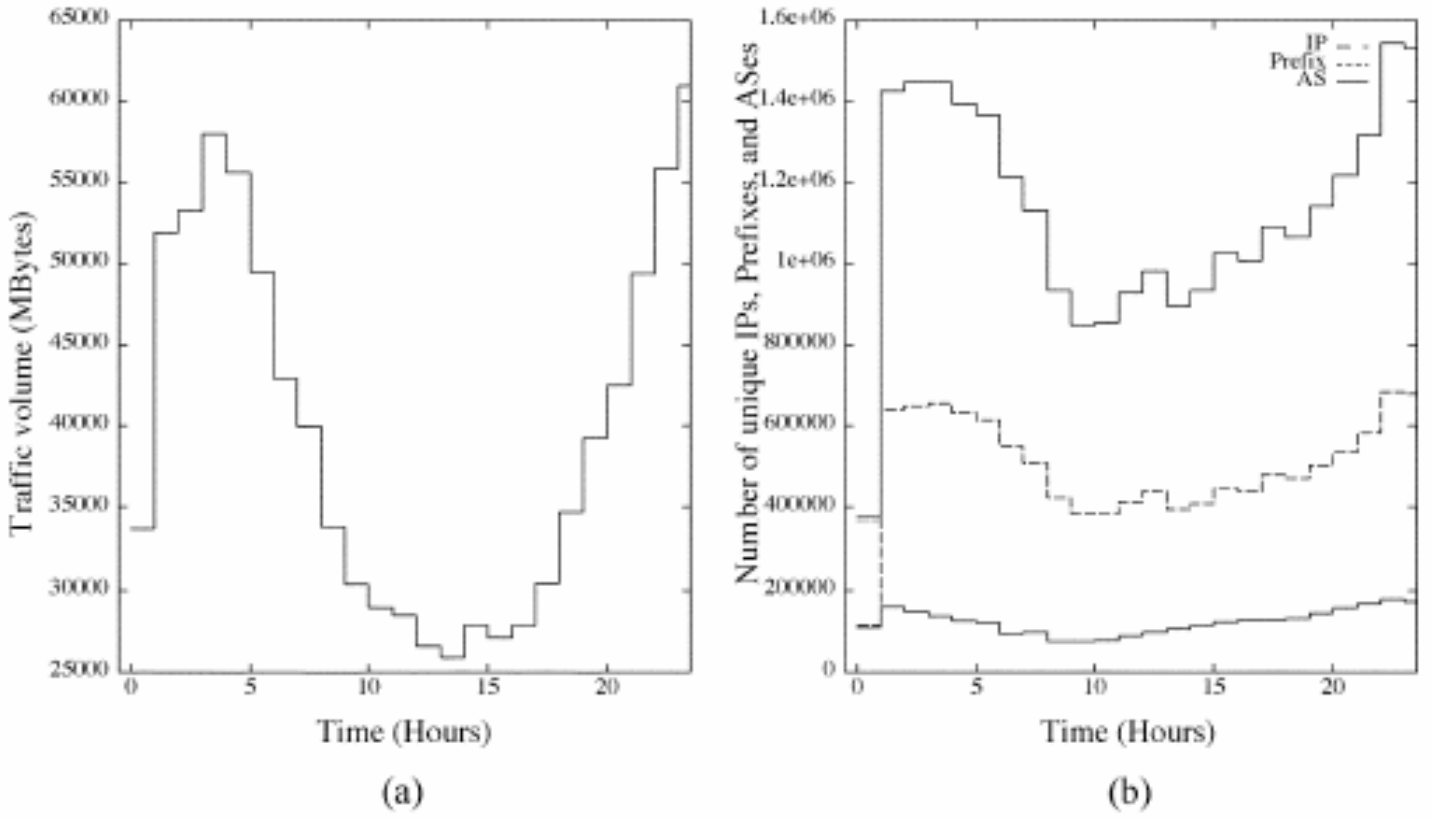
Overview of P2P Traffic and System Dynamics（Cont.）

- Host Connectivity



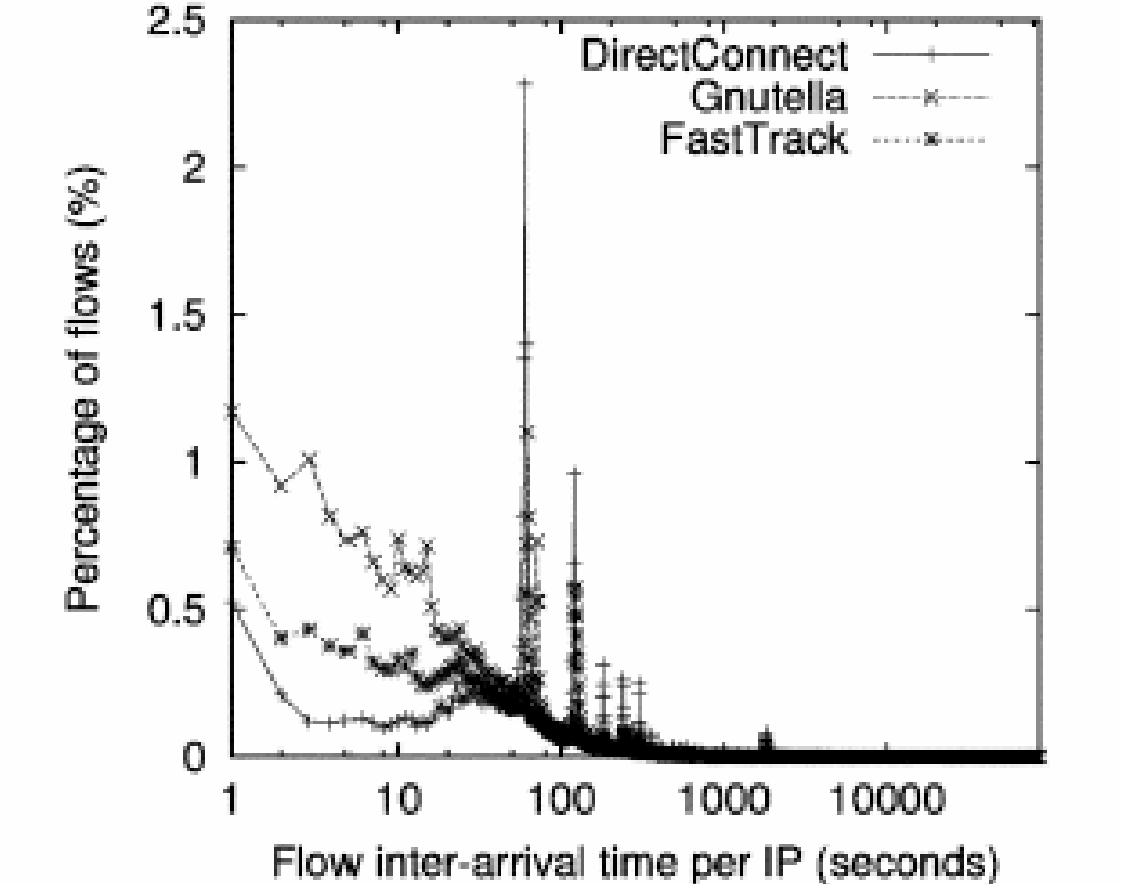
Overview of P2P Traffic and System Dynamics（Cont.）

- Traffic Pattern Over Time



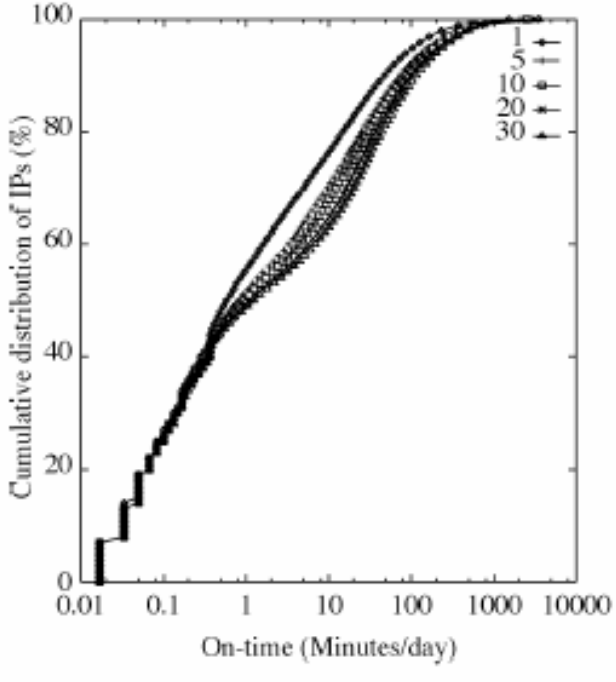
Overview of P2P Traffic and System Dynamics（Cont.）

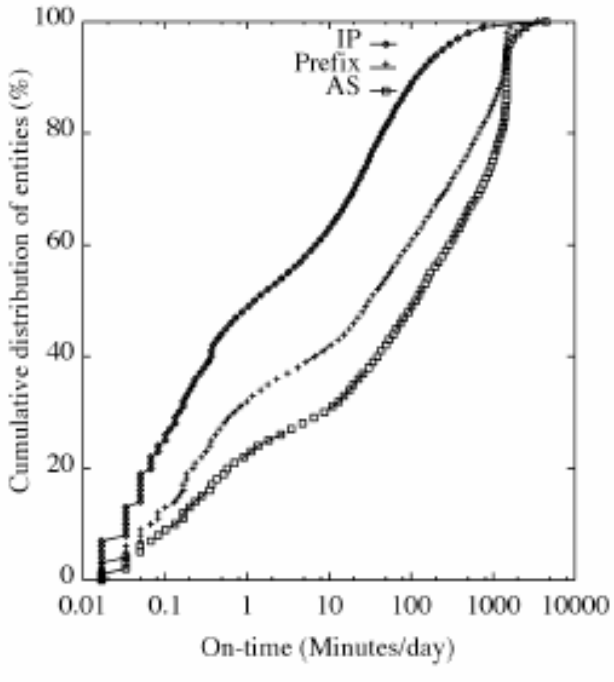
- Host Connection Duration and On-Time

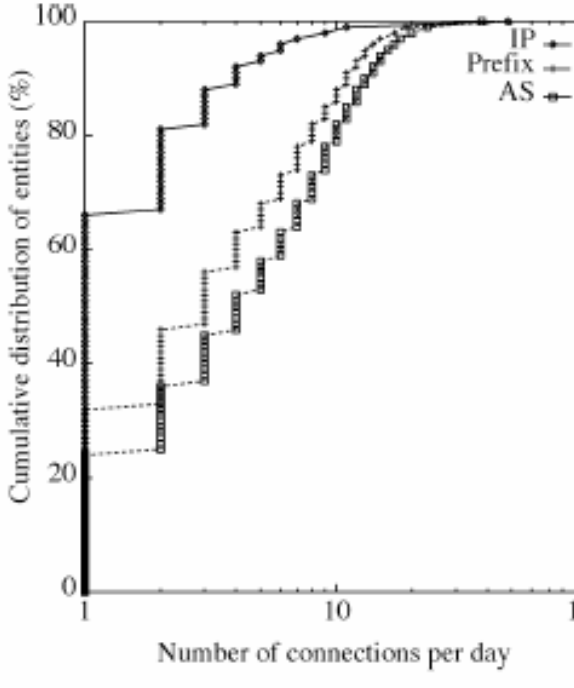


Overview of P2P Traffic and System Dynamics（Cont.）

- Host Connection Duration and On-Time

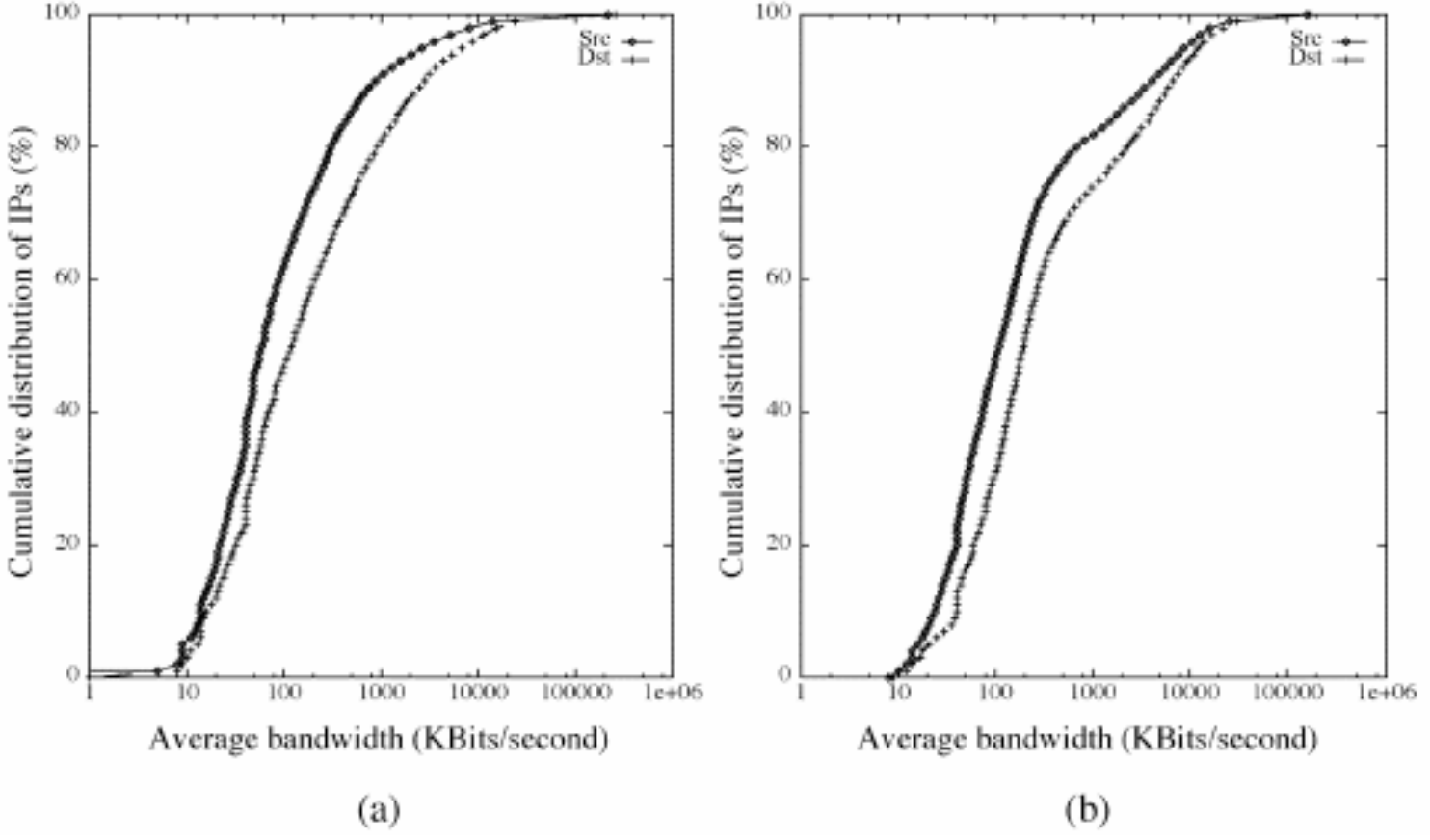






Overview of P2P Traffic and System Dynamics（Cont.）

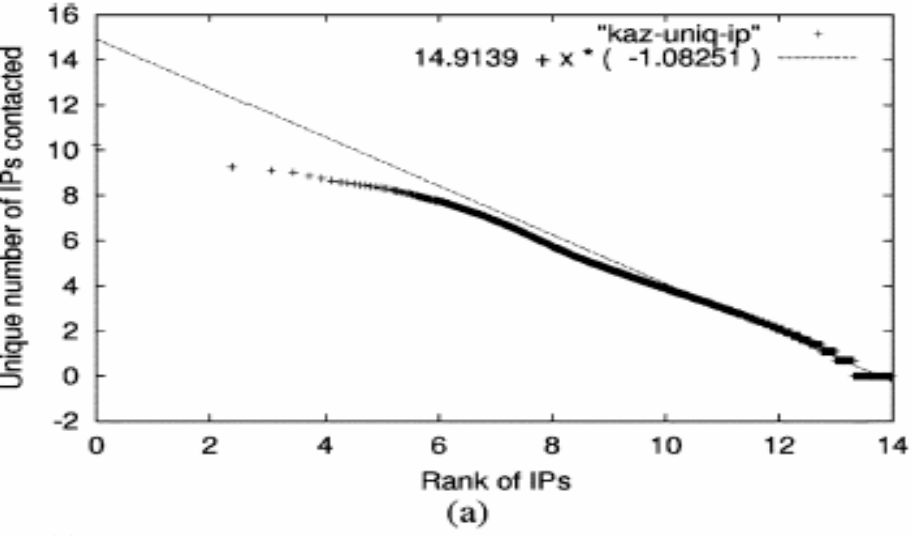
- Mean Bandwidth Usage for Hosts



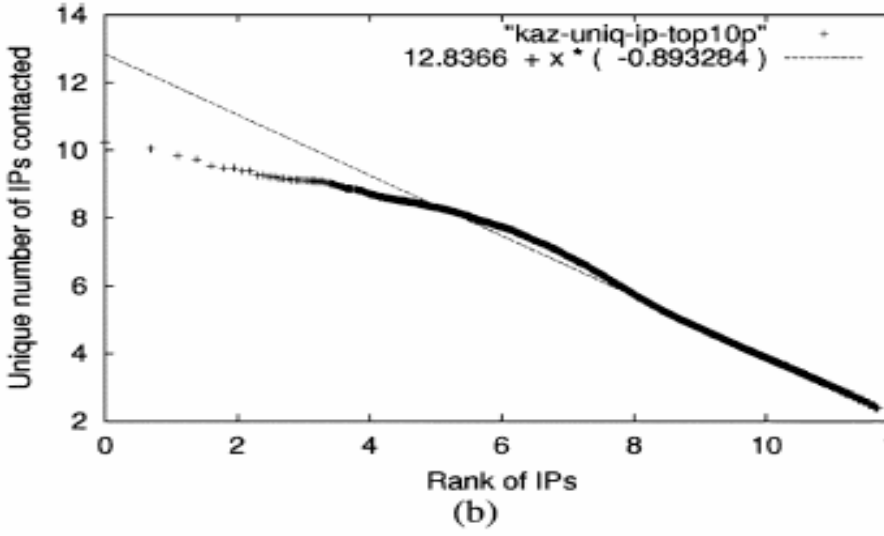
Traffic Characterization

- Zipf's Law
  - The *rank-frequency* plot is the plot of the occurrence frequency f_r of each object versus its rank r, in log-log scales.
  - The rank frequency version of Zipf's law states that $f_r \propto 1/r$.
  - The *generalized Zipf distribution* is defined as $f_r \propto 1/r^{\theta}$, where the slope $-\theta$ is log-log scale can be different than -1.

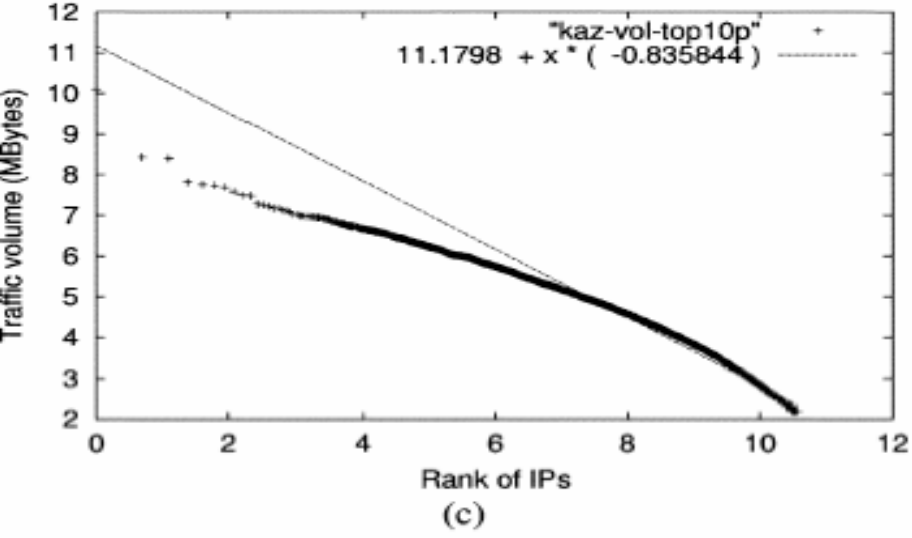
Traffic Characterization (Cont.)



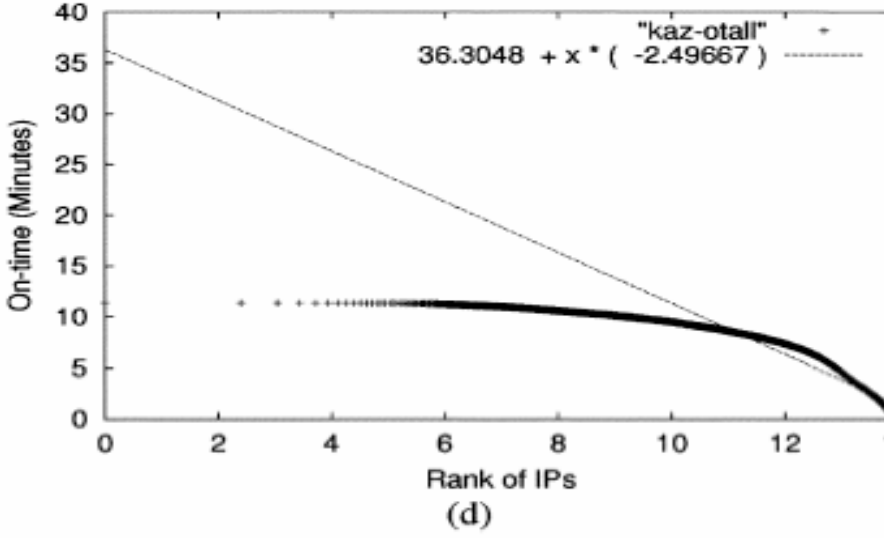
(a)



(b)



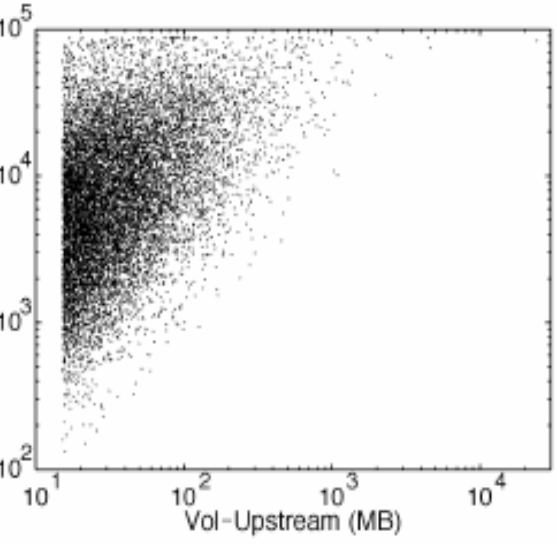
(c)



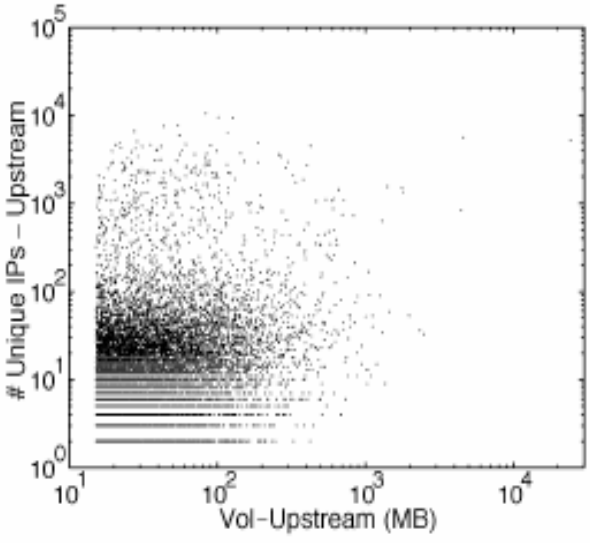
(d)

Traffic Characterization (Cont.)

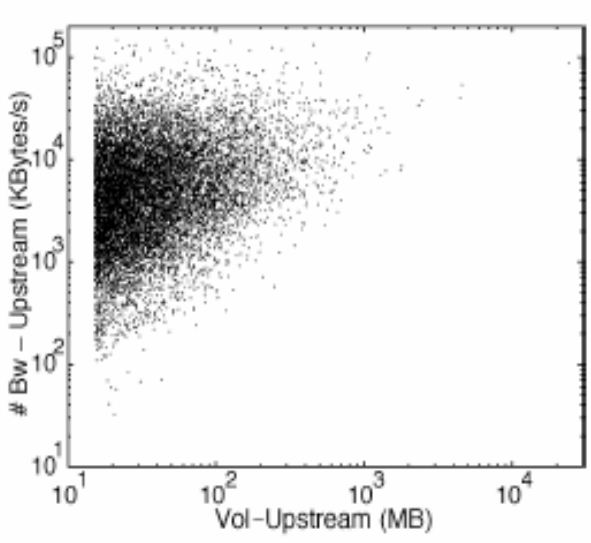
- Relationships Between Measures



(a)



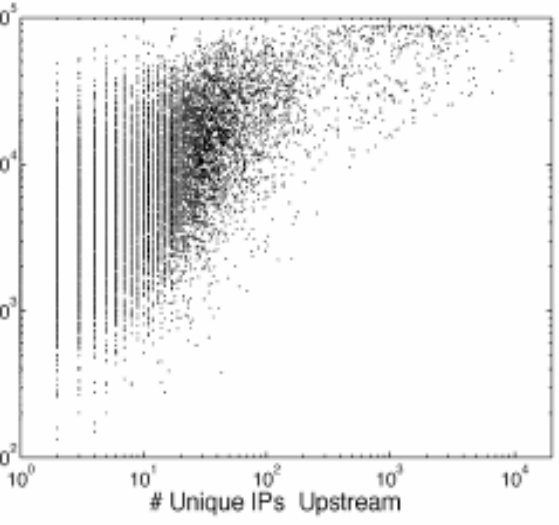
(b)



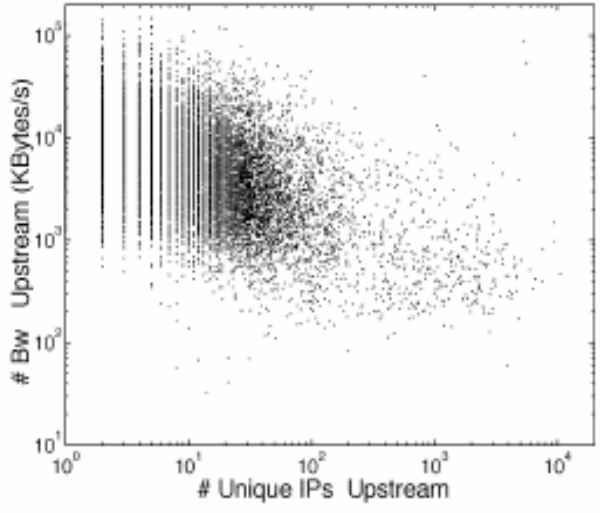
(c)

Traffic Characterization (Cont.)

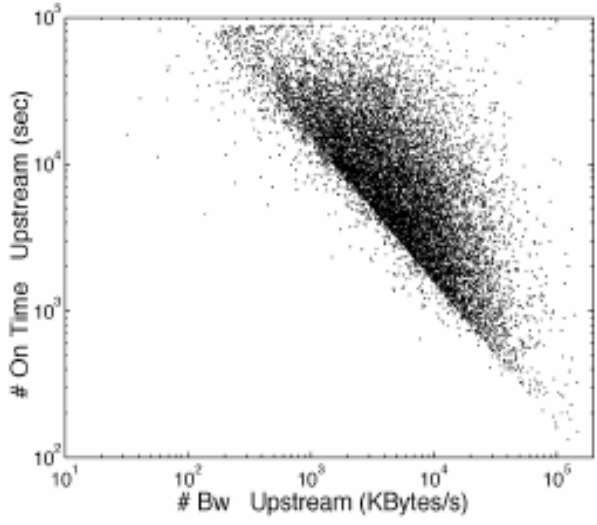
- Relationships Between Measures



(a)



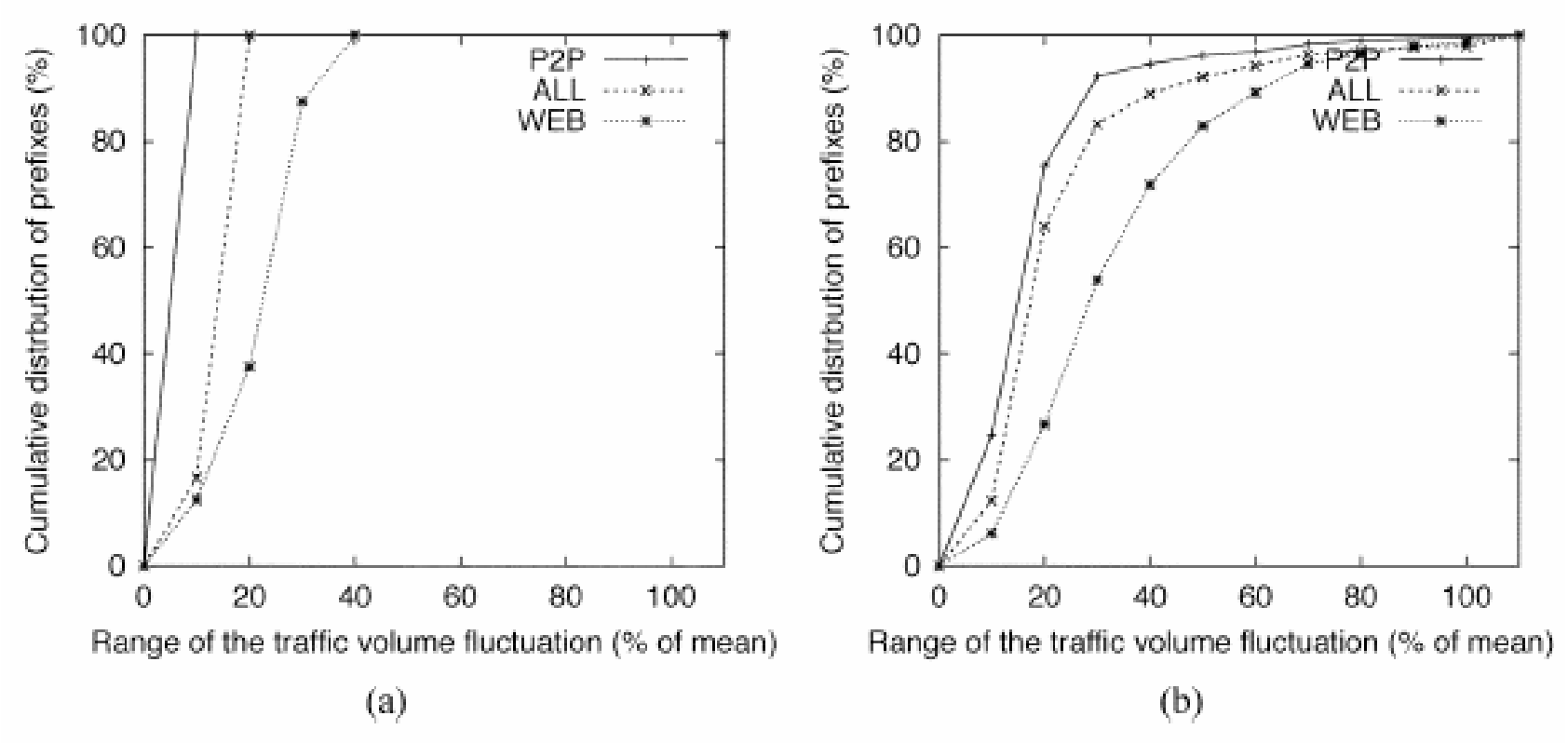
(b)



(c)

Traffic Characterization (Cont.)

- P2P Traffic Versus Web Traffic



Conclusions

- Less than 10% of the IPs contribute around 90% of the total traffic volume.
- The P2P traffic distributions for traffic volume, connectivity, on-time, and average bandwidth usage are extremely skewed; but they don’t fit well with power law distribution.

Conclusions（Cont.）

- All three P2P systems exhibit a high level of system dynamics.
- The fraction of P2P traffic contributed by each network prefix remains relatively unchanged more stable than the distribution of either Web traffic or overall traffic over the time period of one month.

Future Work

- Determine accurate estimates of the peak bandwidth usage
- Develop accurate models for the distributions of the traffic metrics