Understanding Pollution Dynamics in P2P File Sharing

Uichin Lee, Min Choi*, Junghoo Cho
M. Y. Sanadidi, Mario Gerla
UCLA, KAIST*

IPTPS'06

NETWORK RESEARCH LAB. Computer Science Dept. UCLA

Outline

- Pollution in P2P file sharing
- User behavior study
- Pollution model
- Impact on P2P traffic loads
- Conclusion

Pollution in P2P File Sharing

- Pollution is a defensive mechanism to discourage illegal downloads of copyrighted materials.
- Polluting a title (e.g., Maroon5 – This Love)
  - Polluters *aggressively*
    - Tamper content or meta-data of files to create "**polluted versions.**"
    - Pour many polluted versions into the system.
  - Users *powerlessly*
    - Encounter the polluted files with the genuine ones
    - Randomly select one without knowing pollution.

Too Many Polluted Files & Why?

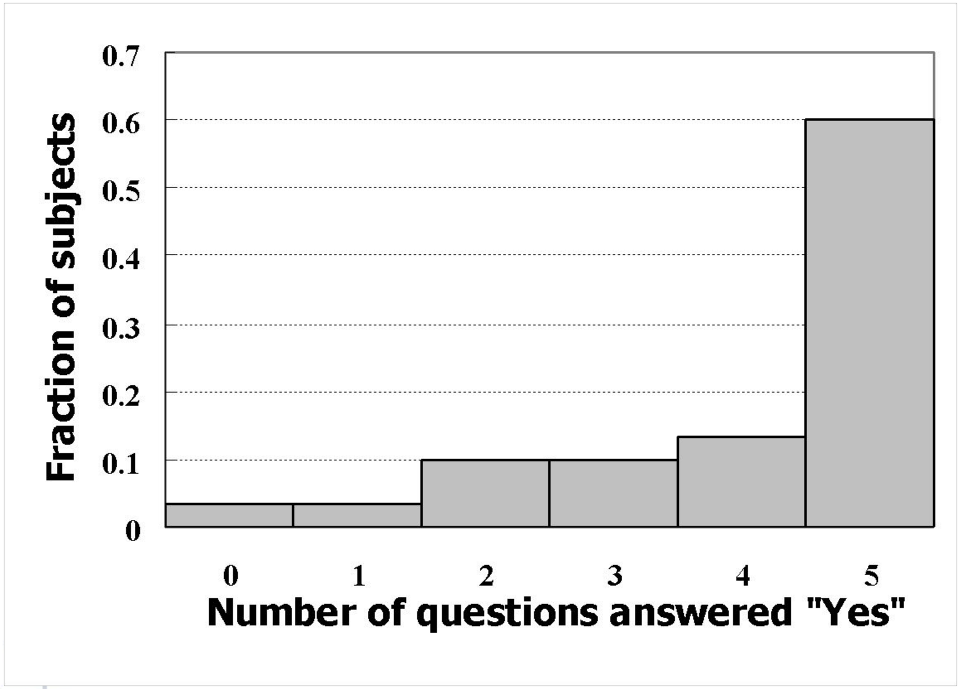
- KaZaA is severely polluted!!
  - Reported by Liang et al. (May, 2004)
  - "My Band" 70% out of 1,816,663 copies
- Given that polluters have limited capabilities (bandwidth/processing power), current level of pollution is too high.
- Recent models were not able to clearly explain such pollution dynamics.
- To better understand pollution dynamics, P2P user behavior must be examined.

User Behavior Study

- Goal
  - How does user behavior impact pollution spread?
- User behavior study setup
  - Two stage test
    - Questionnaire : familiarity / usage patterns
    - Behavior observation : awareness / slackness
  - 30 graduate students (UCLA, KAIST)

Questionnaire
- Familiarity of Participants

- Asked five questions related to P2P
  - Do you know how popular P2P software works?
  - Do you know multi-part downloading or swarming?
  - …



Questionnaire
- Usage Patterns

- P2P usage pattern
  1. Download preparation: send queries/start downloading
  2. Download: check download status
  3. Post-download: check files/decide to share?
- Asked few questions related to each stage

Questionnaire
- Usage Patterns (Results)

1) Preparation Stage

Download decision	57%	quality
	20%	availability
	20%	file size

2) Download Stage

Checking frequency	41%	frequent
	35%	size-dependent
	20%	check later

3) Post-download Stage

Pollution experience	70%	yes
Failed in noticing pollution	30%	yes
Re-download?	23%	yes
	57%	file size dep.

User Behavior Observation

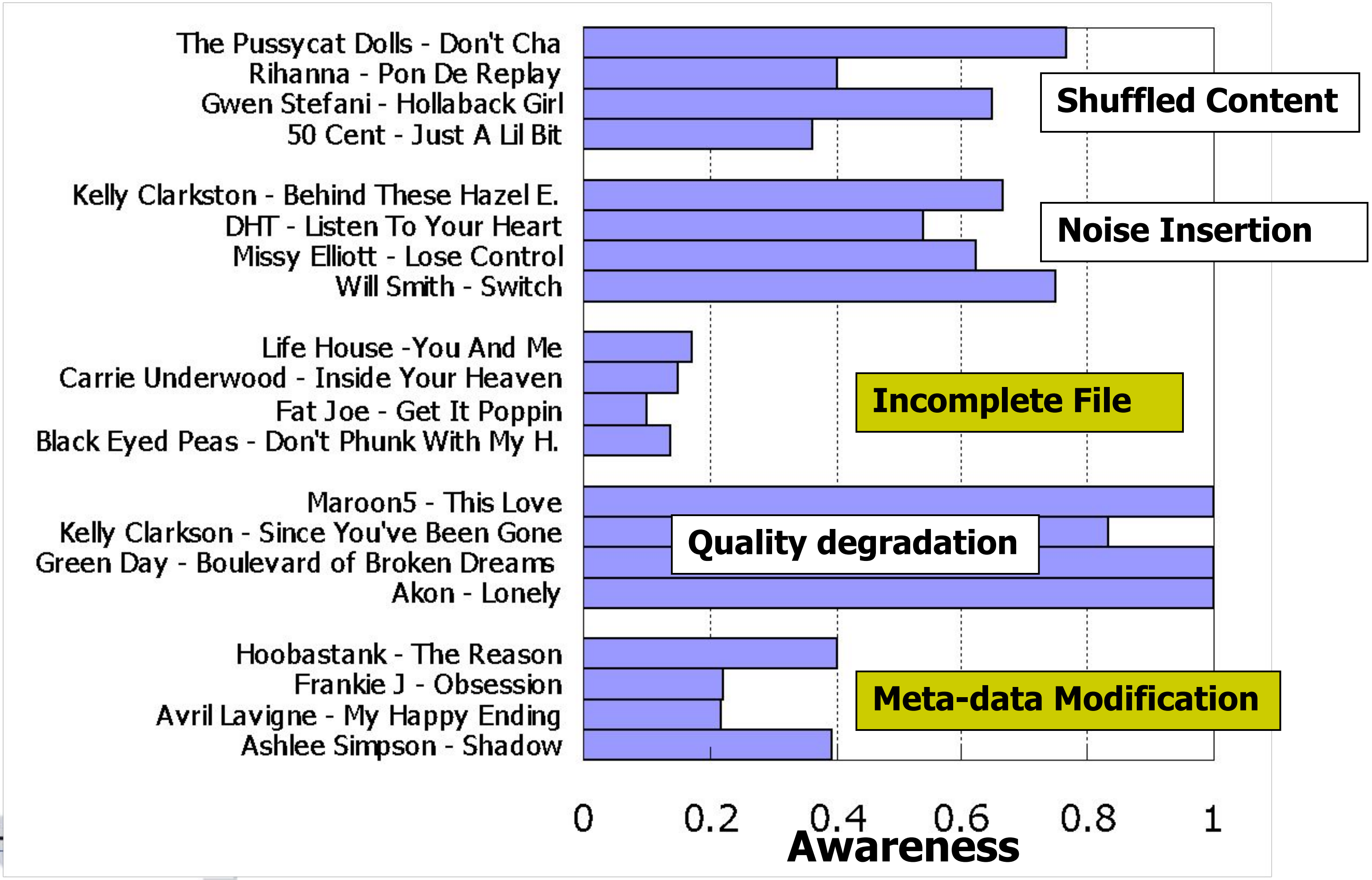
- Metrics
  - Awareness probability : the fraction of users who recognize pollution in a downloaded file
  - Slackness distribution : distribution of intervals between download completion time and quality checking time
- Setup
  - Modified P2P software to monitor user behavior
  - Users are asked to use it and to download files
  - Controlled downloading speed (50K - 1Mbps)

User Behavior Observation

- Pollution techniques (on MP3 files)
  - Meta-data modification : changed names
  - Quality degradation
  - Incomplete file : cut (30-60 seconds beg./end.)
  - Noise insertion: every 15 seconds
  - Shuffled content : randomly shuffled content
- Tested files
  - Applied each pollution technique to four songs
  - 40 popular songs (20 polluted + 20 clean)
- For each download, a user checks familiarity/pollution

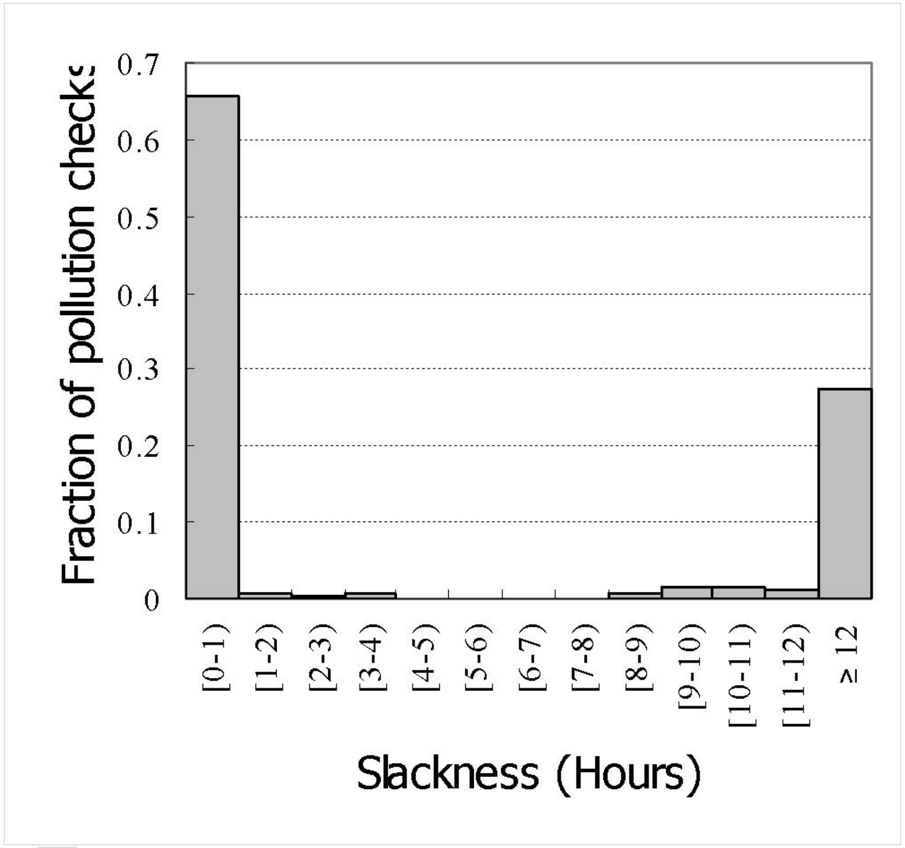
User Behavior Observation

- Awareness



User Behavior Observation

- Slackness – bimodal distribution



Pollution Model

- Discrete time analysis by extending the previous model and incorporating study results
- Total **M** users in the system
- $\mathbf{G_0}/\mathbf{B_0}$: initial # **g**enuine/**b**ogus copies
- Download process

1. At step **k,** a user (never downloaded before) downloads a file with probability $\mathbf{p_s}$ (i.e., interest level)
2. After download, the authenticity is checked after an interval *k* with probability s_k where t <= **L** (max. slackness)
3. Realizes bogus with probability $\mathbf{p_a}$ (i.e., awareness); if so, he will try again with probability $\mathbf{p_r}$ (i.e., re-download prob.)
4. Share the file with probability $\mathbf{p_c}$ (i.e., cooperativeness)

Pollution Model

- # downloads at *k* (N_k)

$$N_k = \underbrace{(M - D_k)p_s}_{\text{New Trials}} + \underbrace{r_k}_{\text{Re-downloads}}$$

M: total number of users
p_s: user's interest rate
D_k: ever downloaded a file
G_k: total # good files
B_k: total # bogus files
r_k: # of re-downloads at k

  - Ever downloaded users (D_k)

$$D_{k+1} = D_k + (M - D_k)p_s$$

- # incoming genuine/bogus files at *k* (g_k , b_k)
  - Total G_k and B_k files are shared in the system

$$g_k = N_k \frac{G_k}{G_k + B_k} \qquad b_k = N_k \frac{B_k}{G_k + B_k}$$

Pollution Model

- Total # genuine files (G_{k+1})

$$G_{k+1} = G_k + g_k - (1-p_c)\sum_{j=1}^{L} s_j g_{k+1-j}$$

- Total # bogus files (B_{k+1})

$$B_{k+1} = B_k + b_k - p_D \sum_{j=1}^{L} s_j b_{k+1-j}$$

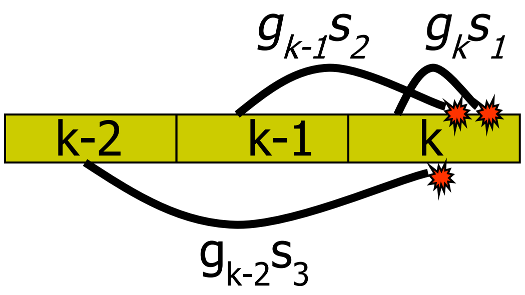
  - Prob. of not sharing bad files

$$p_D = p_a + (1-p_a)(1-p_c)$$

- # re-downloads at *k+1*

$$r_{k+1} = p_r p_a \sum_{j=1}^{L} s_j b_{k+1-j}$$

Max Slackness: L=3



Total # checking at *k*:
$g_{k-2}s_3 + g_{k-1}s_2 + g_k s_1$

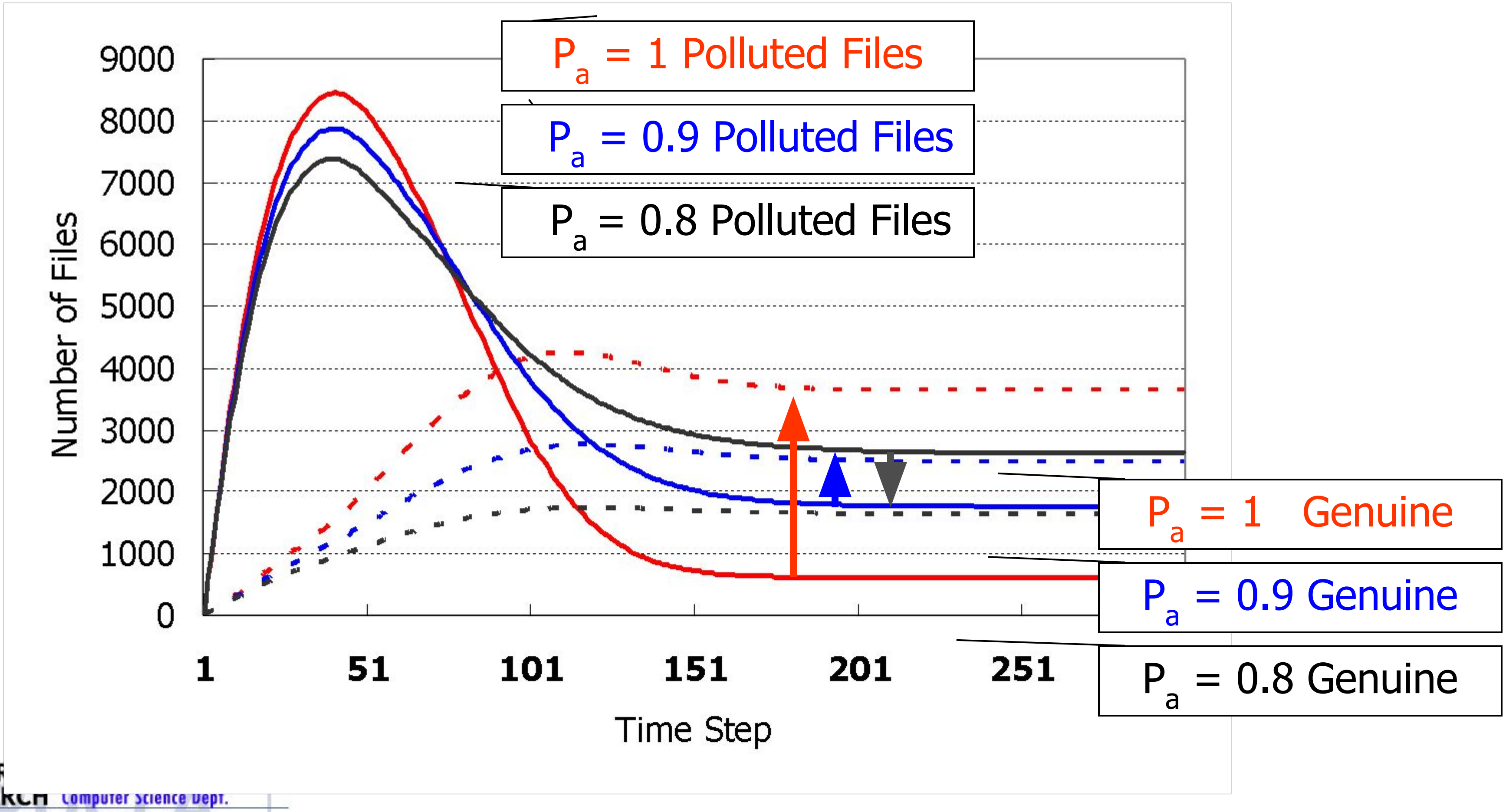
L: max. slackness
g_k: # incoming good files at k
b_k: # incoming bad files at k
s_j: prob. of checking after j
p_c: cooperation probability
p_a: awareness probability
p_r: re-download probability

Pollution Model

- Analysis
  - Iterative solution of the proposed model
  - Metric
    - Pollution level = # polluted copies / # genuine copies
  - Setting
    - M=15,000 (total number of users)
    - L=48 (max. slackness)
    - p_s = 1/24 (gets interested in every 24 hrs.)
    - p_r = 1 (re-downloads always!)
    - p_c = 0.25 (cooperativeness)
    - Initial pollution level = 20

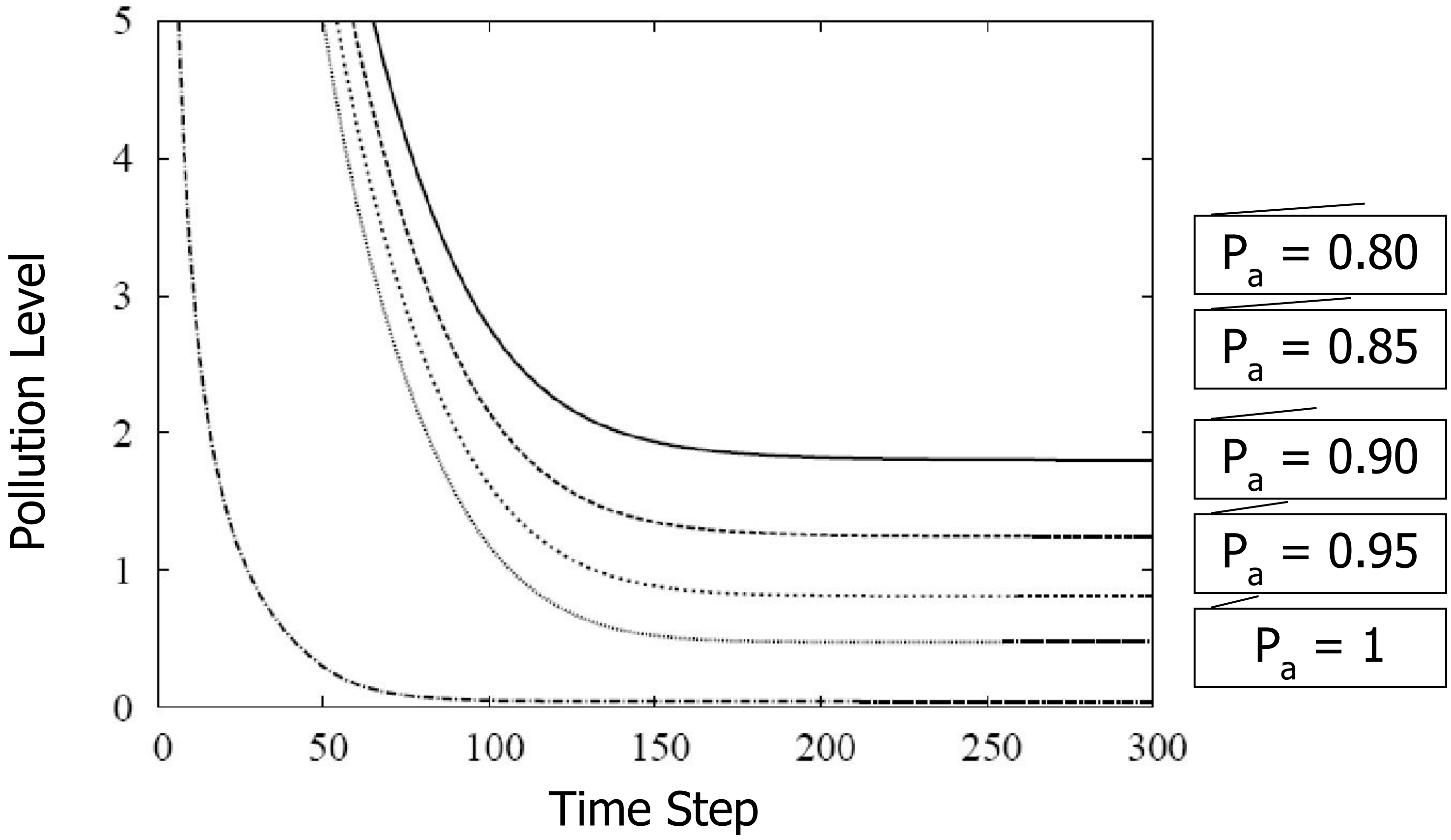
Pollution Model

- Impact of user awareness



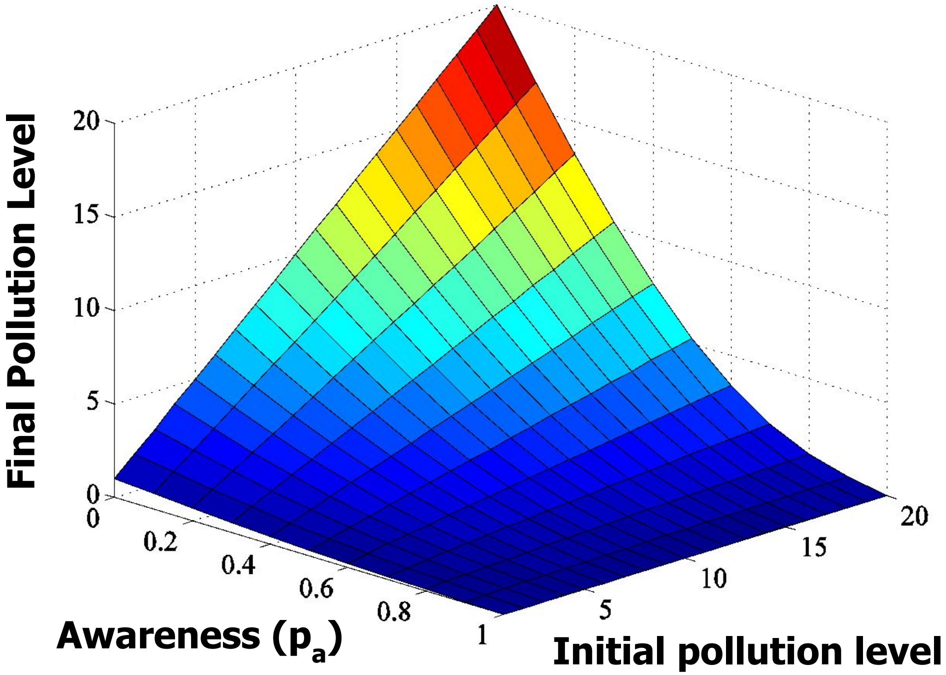
Pollution Model

- Pollution level (k) = #polluted (k)/#genuine (k)



Pollution Model

- Increasing initial pollution level
  - Awareness is critical to make an effective attack



Impact on P2P Traffic Loads

- Popular files are targets of the polluters!!
- Users will re-download with probability p_r
- From our model, we can estimate the total # of re-downloads
- In the worst case, # of re-downloads is **x3** larger!!
- 60% of the Internet traffic is P2P

Conclusion

- User behavior study shows
  - Users are not error-free in recognizing pollution
  - Users' slackness follows a bimodal distribution
- Developed an analytical model
- Analytic model shows
  - Awareness is one of the key factors in pollution dynamics
  - Pollution has a great impact on the P2P traffic loads