CCNC 2009

A Collusion-Resistant Automation Scheme for Social Moderation Systems

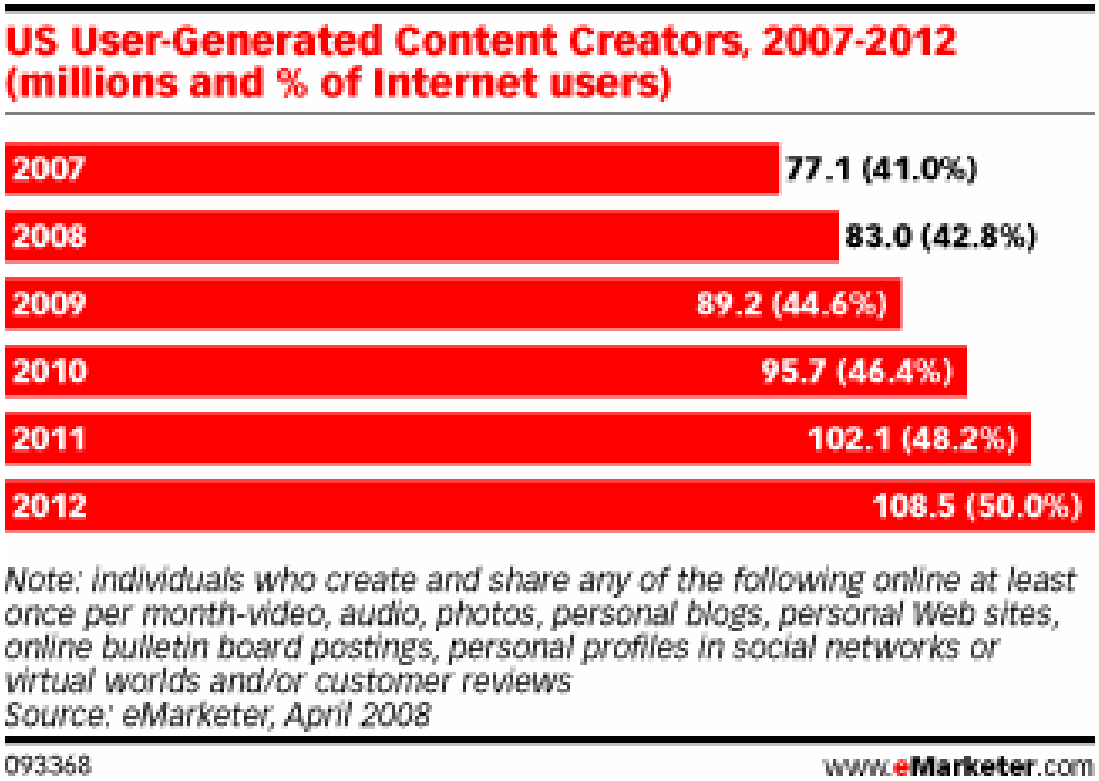
Jing-Kai Lou[12], Kuan-Ta Chen[1], Chin-Laung Lei[2]

Institute of Information Science, Academia Sinica[1]

Department of Electrical Engineering, National Taiwan University[2]

The Rise of User-Generated Content

- eMarketer projects that the number of US UGC creators will rise to **108 million** in 2012, from **77 million** in 2007.



Inappropriate UGC

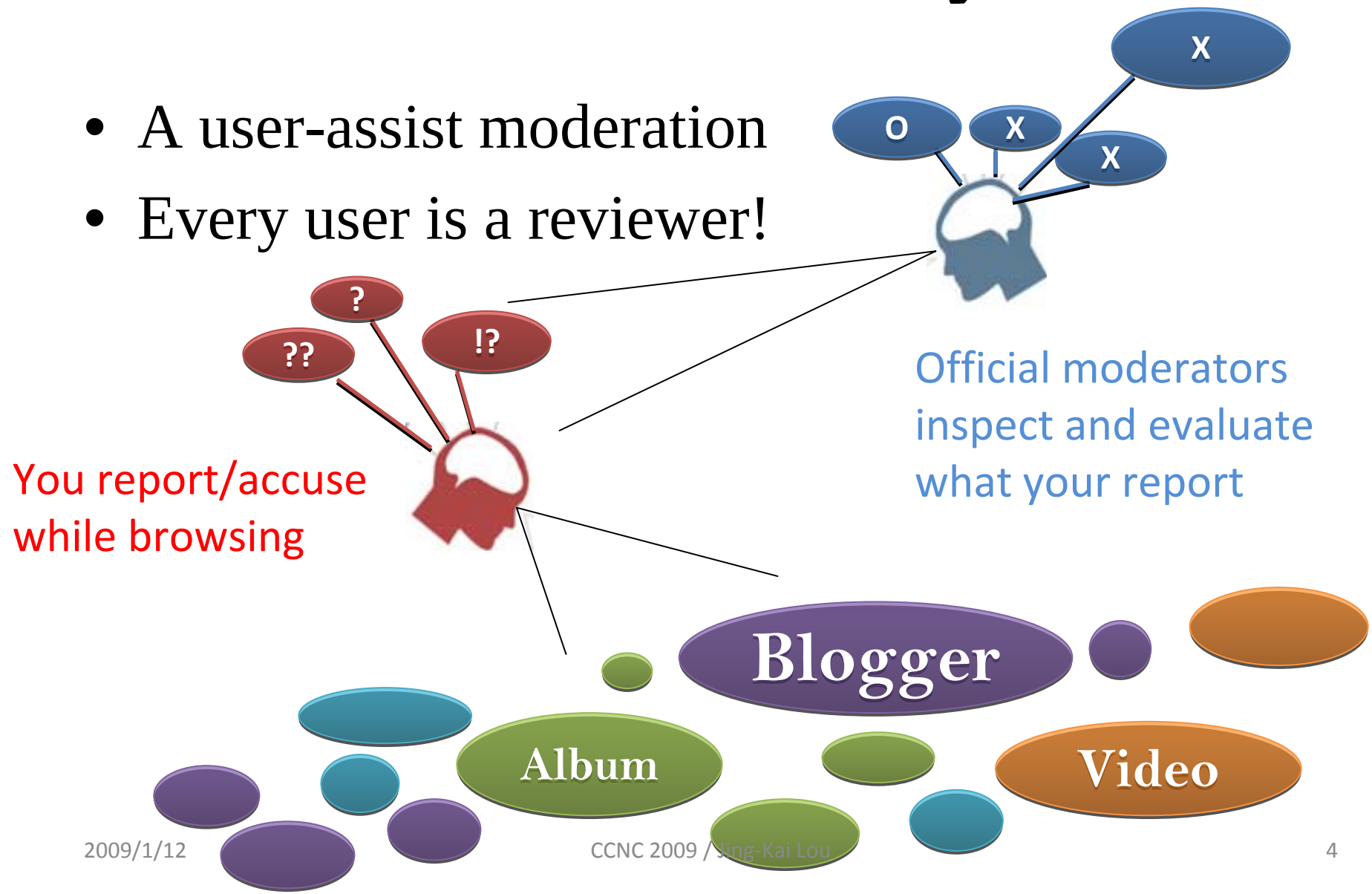
- **While most UGC creators behave responsibly, a minority of creators may upload inappropriate content, such like**
  - pictures that violate copyright laws
  - splatter movies
  - …
- ***Content censorship is essential*** for Web 2.0 services

- **One Solution:**
  - **hiring lots of official moderators**
  - But, such **high labor cost** is a great **burden** to the service provider

- **Another Solution:**
  Social moderation has been proposed to solve the content censorship problem

Social Moderation System

- A user-assist moderation
- Every user is a reviewer!



Is Social Moderation good enough?

- Advantages of social moderation system:
  1. **Fewer official moderators**
  2. **Detecting inappropriate content quickly**

- **BUT**, the number of the reports is still large.
  - Even **1%** uploading photos in Flickr are problematic, there are about **43,200 reports each day**.
- Can we help?

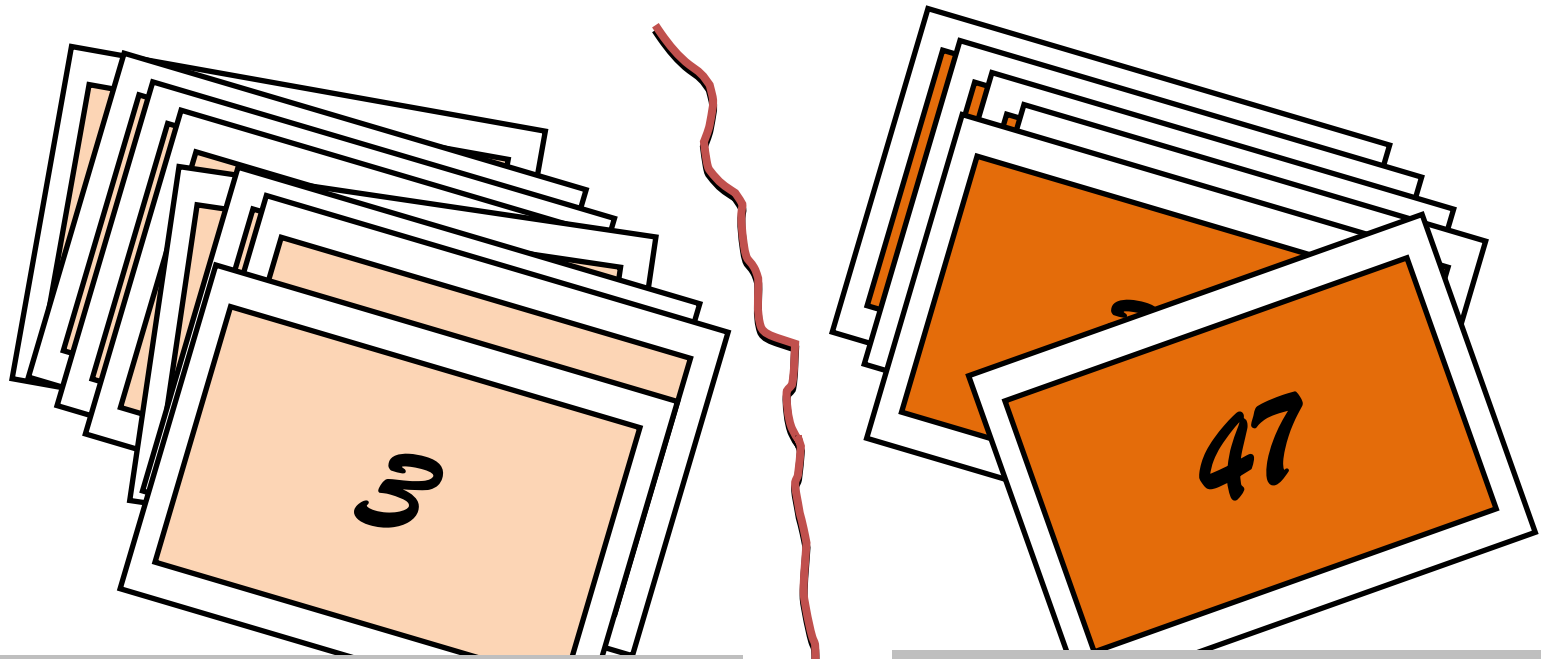
Social Moderation Automation

- This is our motivation for proposing social moderation automation, which automatically summarizes the reports submitted by users.

- **A preprocess:**
  For eliminating manual inspection by official moderators as much as possible.

There is an intuitive way…

- **Count-based Scheme** identifies **misbehaving users** by considering the number of accusations (reports).



These photos are accused no more than ($N=20$) users

These photos are accused more than ($N=20$) users

However, there are many colluders…

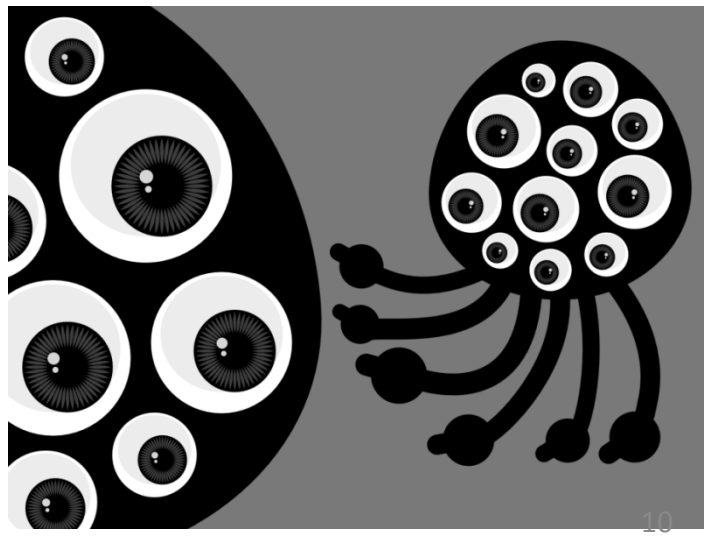
Not All Users Are Trustable

- While most users report responsibly, **colluders report fake results** to gain some benefits.
- **Counted-based scheme may misidentify!**

Research Question

- **CAN** we automatically infers which accusations (reports) are **fair** or **malicious**?

- Need a better automation scheme to deal with collusion attacks



Our Scheme

- **Community-based scheme** analyzes the *accusation relations* between the accusing users and accused users.

- Based on the derived information, the scheme infers whether the accusations are fair or malicious;
  that is, it distinguishes users that genuinely misbehave from victims of collusion attacks.

Our Contributions

- **The evaluation results show that our scheme**
  - Achieves **accuracy rate** higher than **90%**
  - Prevents at least **90%** victims from collusion attacks

Accusation Relation

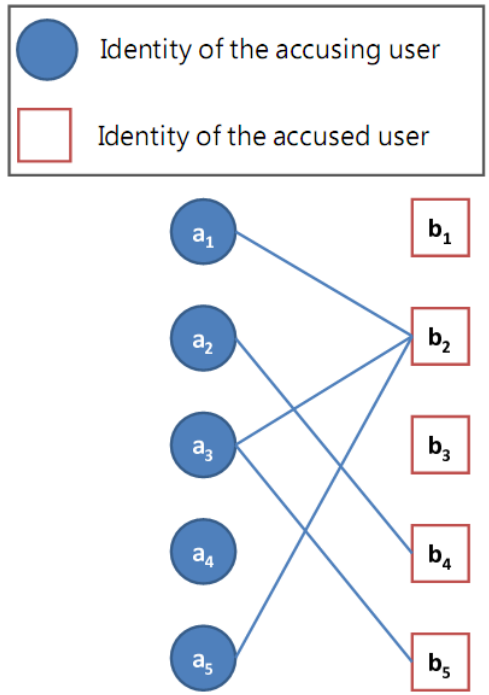
- **Accusation Relation(R):** a subset of A x A,
  A := {reporters, UGC creators}
- E.g. 5 users in this system, namely U1, U2, U3, U4, U5
- **Accusation Relation Matrix(M):**
  - U1 accuses (reports) U2
  - U2 accuses U4
  - U3 accuses U2 & U5
  - U4 accuses none
  - U5 accuses U2

User	1	2	3	4	5
1	0	1	0	0	0
2	0	0	0	1	0
3	0	1	0	0	1
4	0	0	0	0	0
5	0	1	0	0	0

Accusing Graph

- Input for our community-based scheme
- **Accusing Graph(G):**
  - An undirected bipartite graph G(A+B, E)
  - A: {accusing identity of users}
  - B: {accused identity of users}

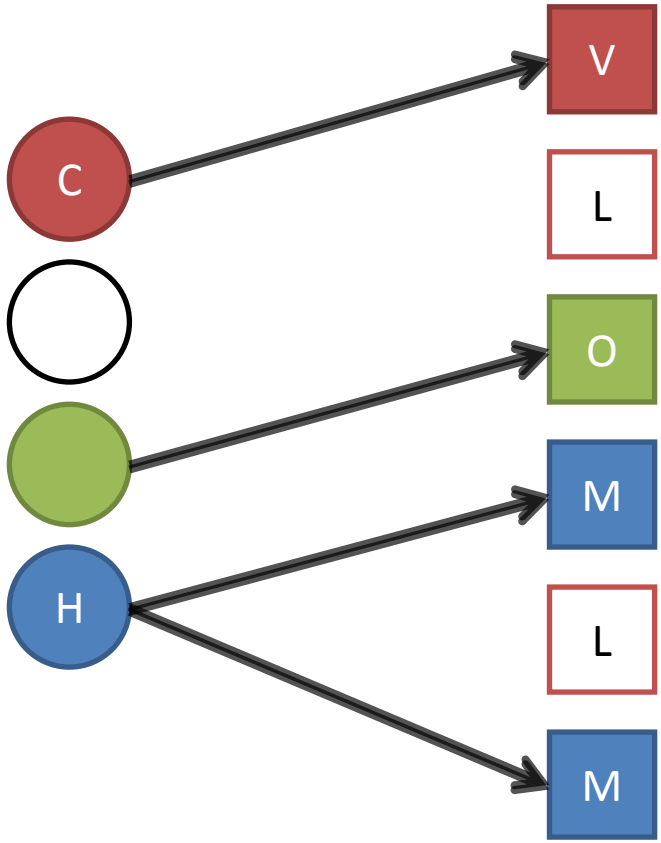
User	1	2	3	4	5
1	0	1	0	0	0
2	0	0	0	1	0
3	0	1	0	0	1
4	0	0	0	0	0
5	0	1	0	0	0



Meanings of Nodes

Identity of accusing user

- **Colluders**
- **Careless accuser**
- **Honest accuser**
- **User doesn't accuse**

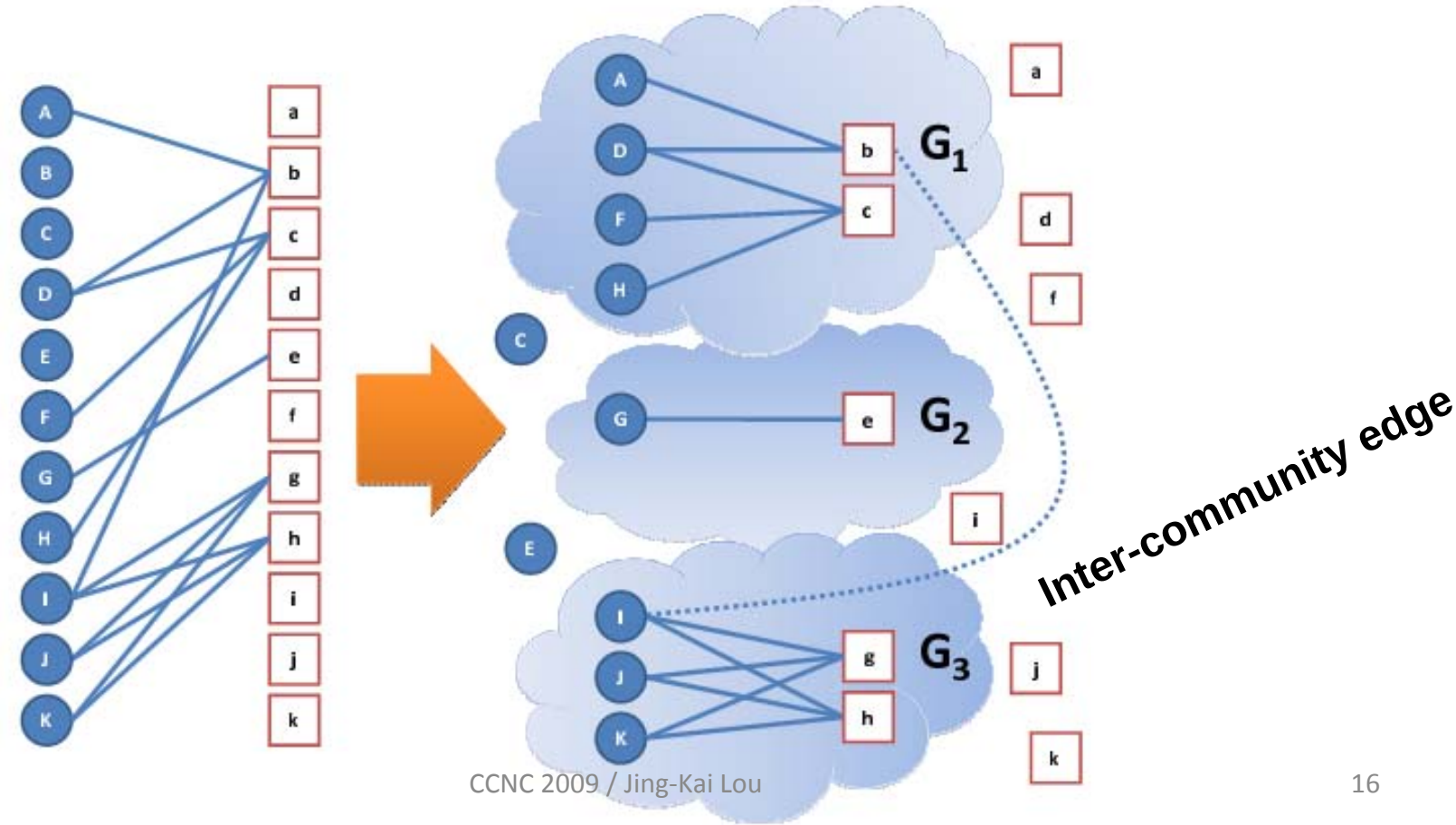


Identity of accused user

- **Victim**
- **Unfortunate user**
- **Misbehaving user**
- **Low-abiding user**

Accusing Community

- Adopting Girvan-Newman Algorithm to detect the **communities** and the **inter-community edges**

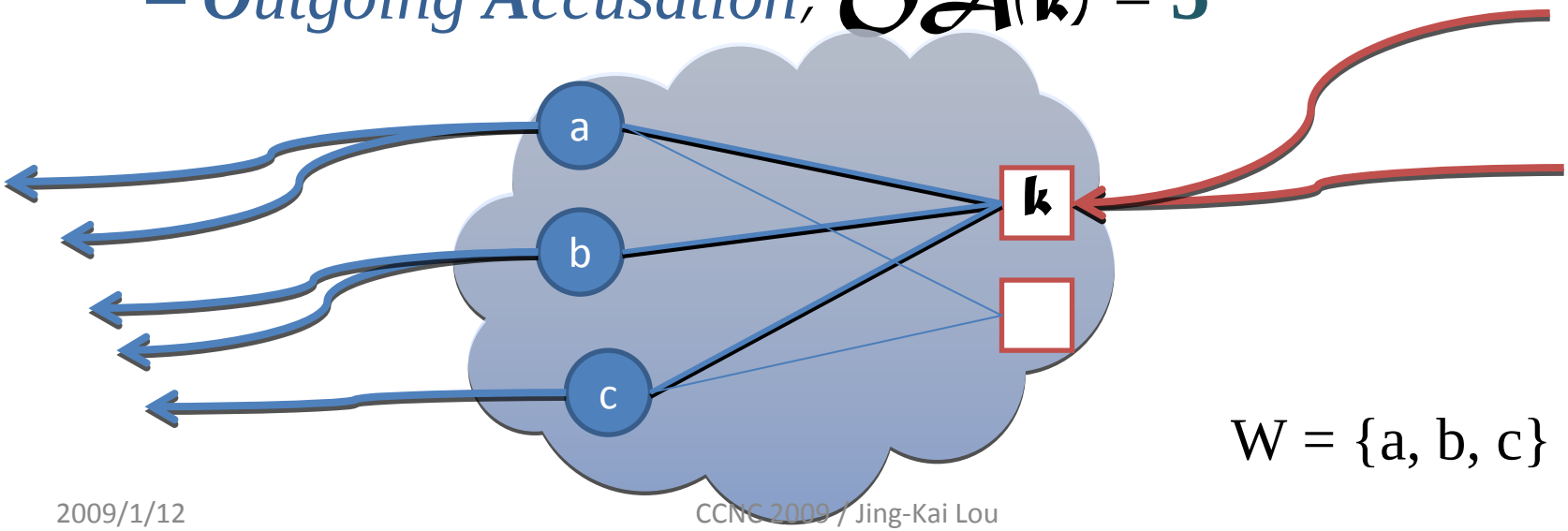


Inter-community edge

- **Property 1:**
  It is unlikely that an inter-community edge is an accusing edge between a **colluder** and a **victim**.
- **Property 2:**
  It is unlikely that an inter-community edge is an accusing edge between a **careless accuser** and an **unfortunate user**.
- **Property 3:**
  An inter-community edge most likely is an accusing edge between an **honest accuser** and a **misbehaving user**.

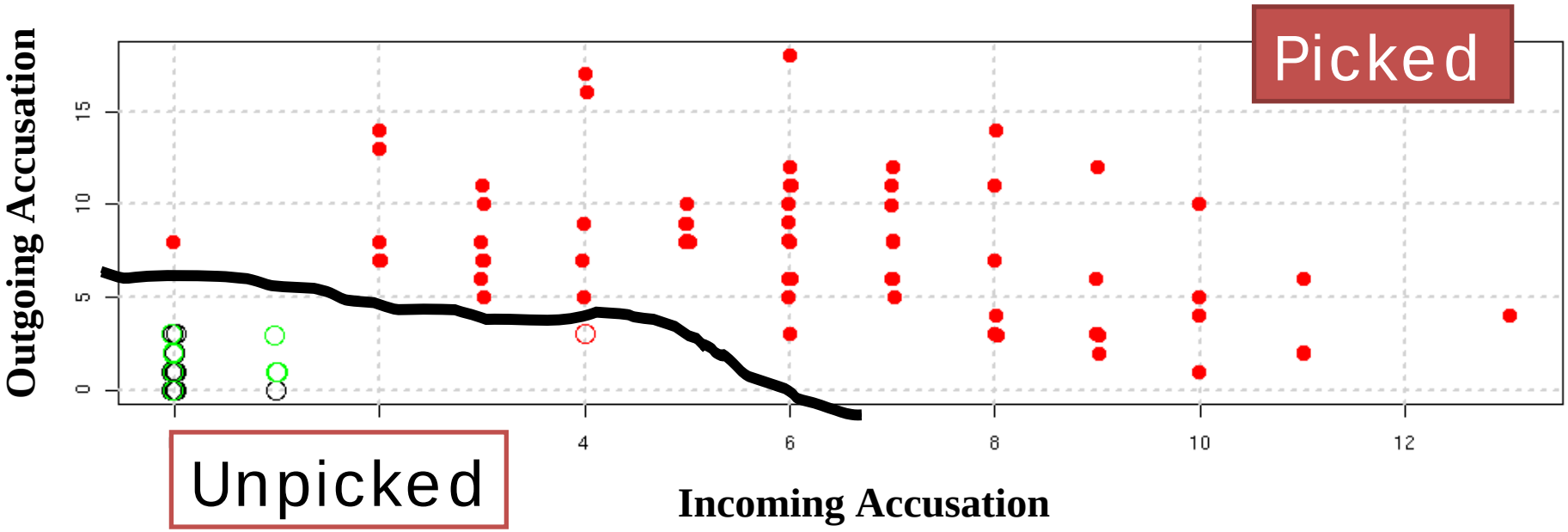
Features for each User

- inter-community edges ➔ fair accusations
- Base on the inter-community edges, we design features for nodes
  - *Incoming Accusation*, $\mathcal{IA}(k) = 2$,
  - *Outgoing Accusation*, $\mathcal{OA}(k) = 5$



$W = \{a, b, c\}$

Clustering (IA, OA) pairs



Unpicked unfortunat users

Unpicked misbehaving users

Unpicked victims

Picked Unfortunate users

Picked misbehaving users

Picked victims

Algorithm

1. Partitioning accusing graph into communities.

2. Computing the feature pair ($\mathcal{JA}$, $\mathcal{OA}$) of each user

3. Clustering based on their ($\mathcal{JA}$, $\mathcal{OA}$) pairs, and label users in the cluster with larger ($\mathcal{JA}$, $\mathcal{OA}$) as misbehaving users.

Simulation Setup

- We use simulations to evaluate the performance of our scheme in detecting real misbehaving users in a social moderation system.

- Simulation Assumption:

1. A **honest user** should only accuses users that definitely **misbehave**.
2. A **colluder** accuses **victims**.
3. All users including colluders have a probability *of making an accusation* ***by mistake.***

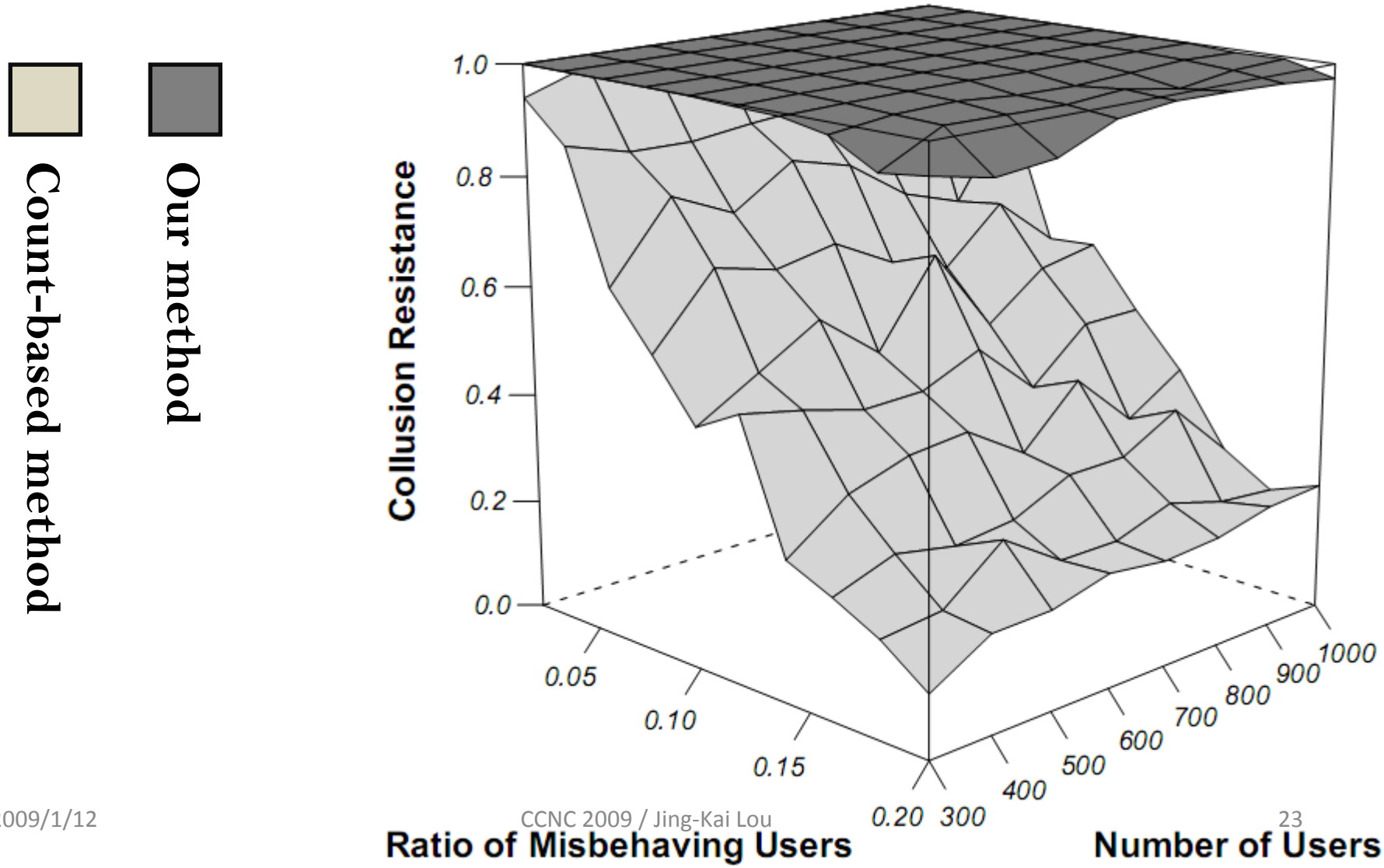
Evaluation Metric

- **What we care is, False Negative**
  - Misidentifying victims as misbehaving users

- **Collusion Resistance**

$$collusion\ resistance = 1 - \frac{|\ misidentified\ victims\ |}{|\ all\ victims\ |}.$$

Effect of #(Misbehaving users)



Conclusion

- We propose a **community-based scheme** based on the **community structure of an accusing graph**.
- The results show that the collusion resistance of our scheme is around **90%**.
- We believe that **collusion-resistant schemes** will play an important role in the design of social moderation systems for Web 2.0 services

Thank you for your listening

Q&A