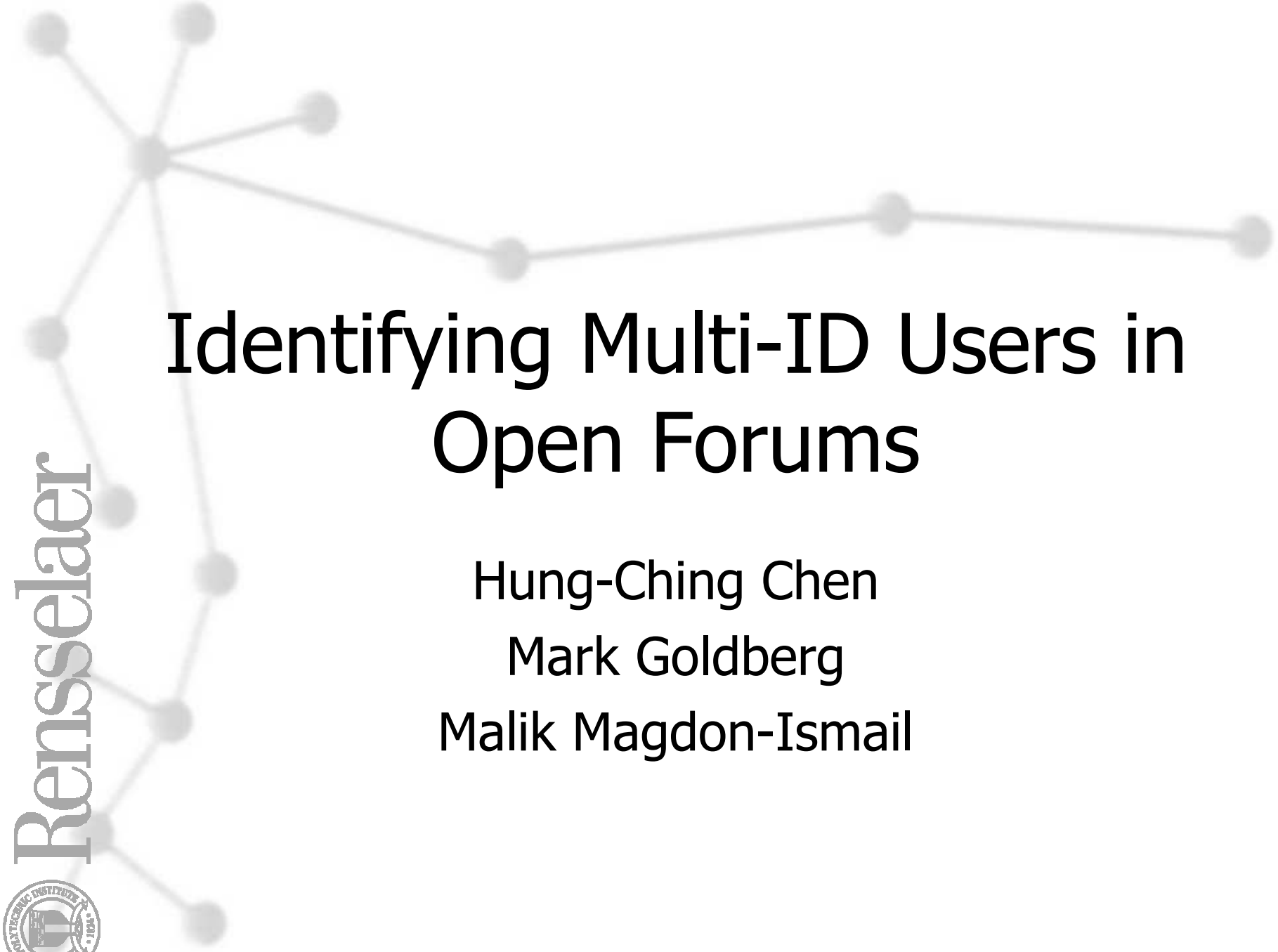


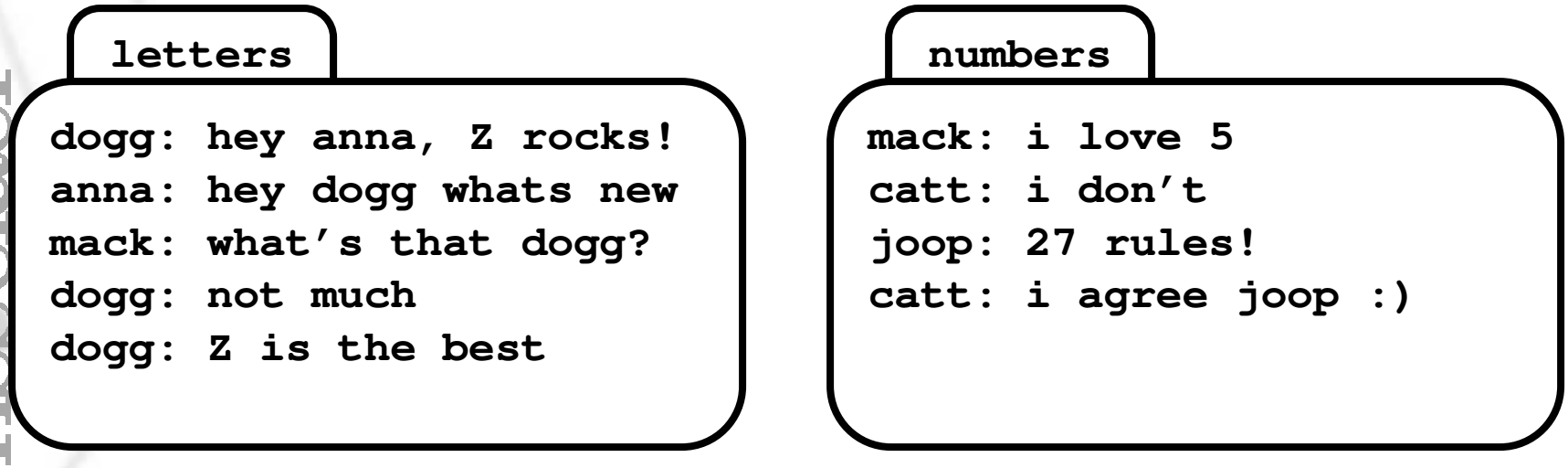
Identifying Multi-ID Users in Open Forums

Hung-Ching Chen
Mark Goldberg
Malik Magdon-Ismail

Rensselaer

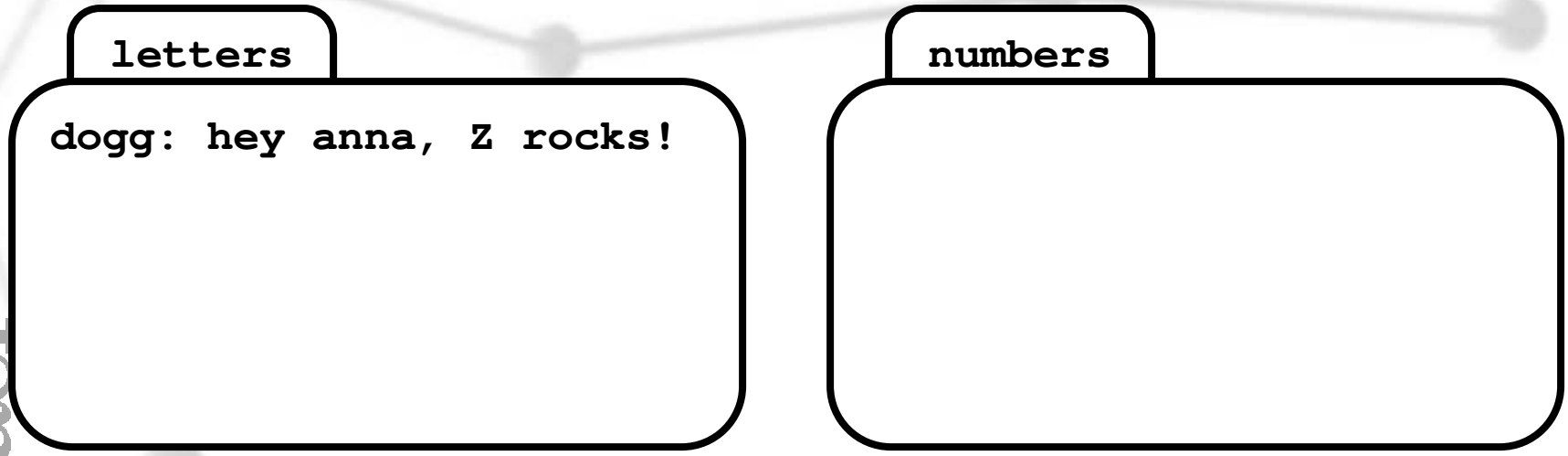
Who is Using Multiple IDs?

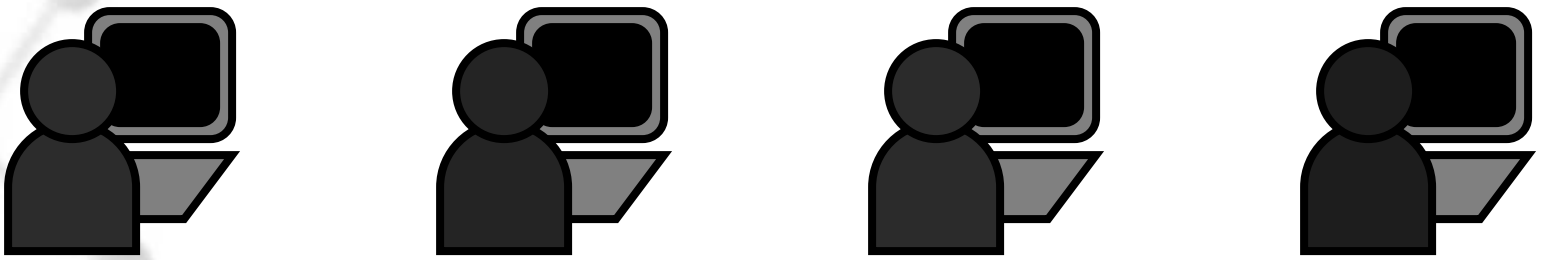
Consider two chatrooms, “letters” and “numbers”



Rensselaer

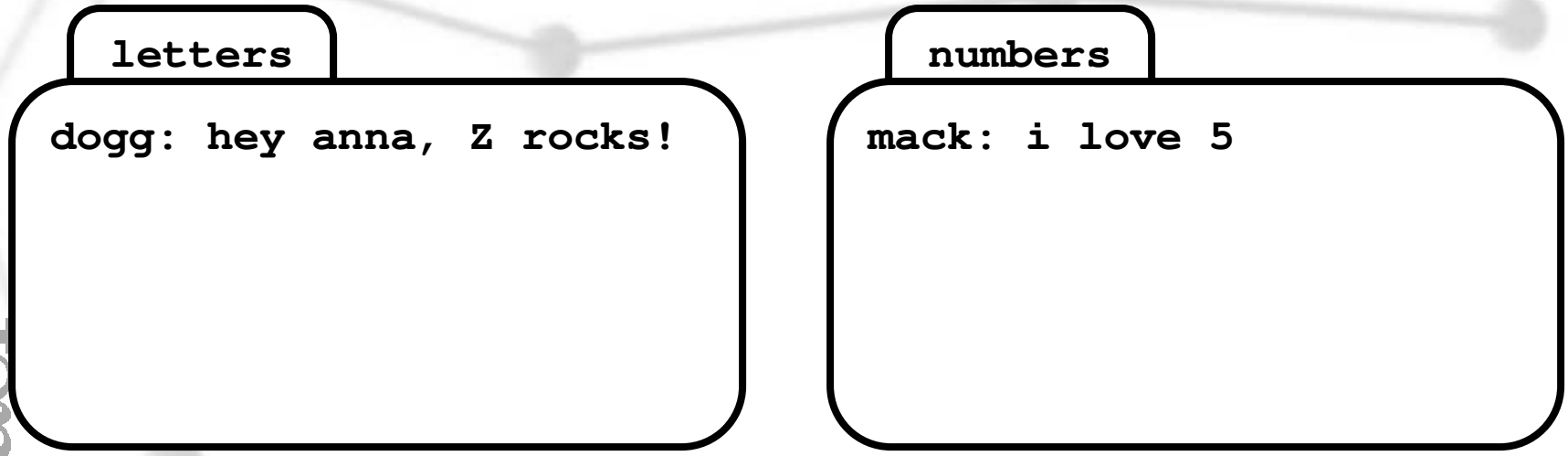
Multi-ID Users

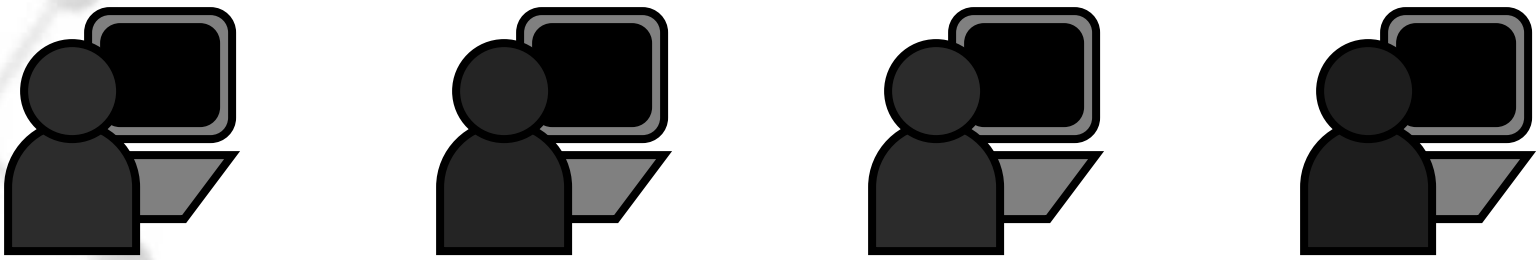




Rensselaer

Multi-ID Users





Rensselaer

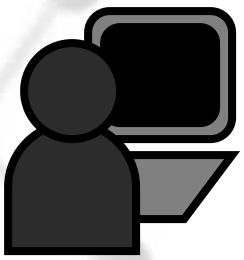
Multi-ID Users

letters

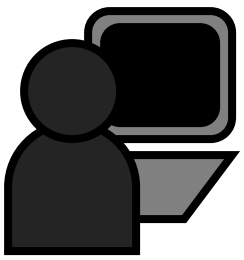
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
```

numbers

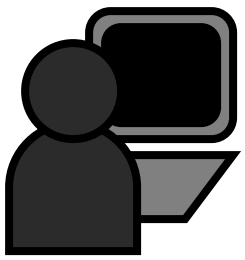
```
mack: i love 5
```



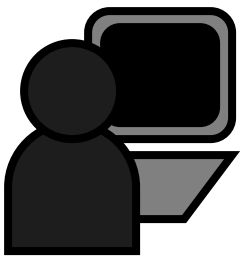
anna



mack



dogg/catt



joop

Rensselaer

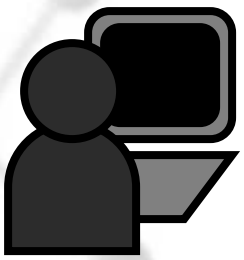
Multi-ID Users

letters

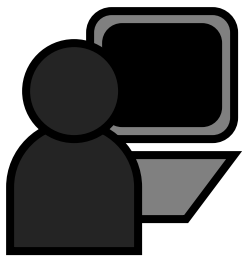
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
```

numbers

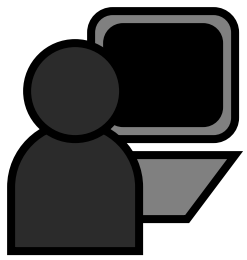
```
mack: i love 5
```



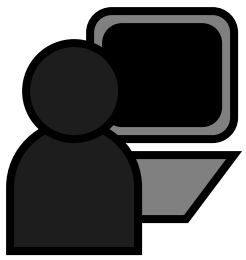
anna



mack



dogg/catt



joop

Rensselaer

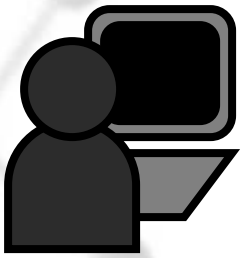
Multi-ID Users

letters

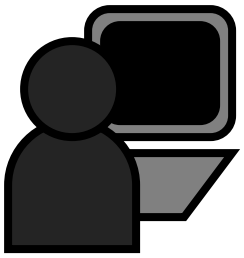
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
```

numbers

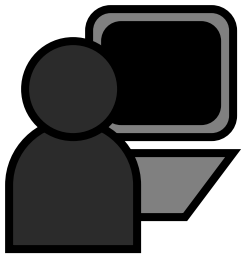
```
mack: i love 5

catt: i don't
```



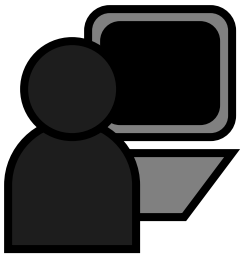
anna



mack



dogg/catt



joop

Rensselaer

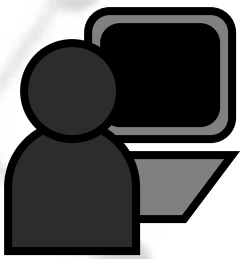
Multi-ID Users

letters

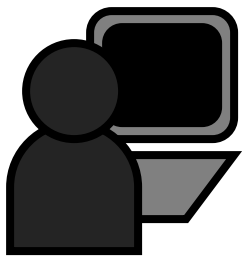
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
```

numbers

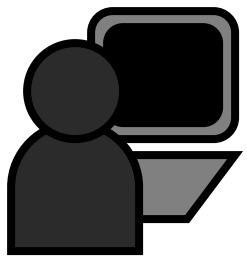
```
mack: i love 5

catt: i don't
```



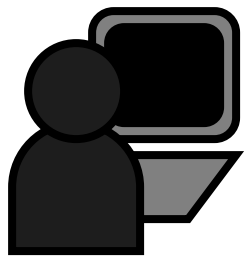
anna



mack



dogg/catt



joop

Rensselaer

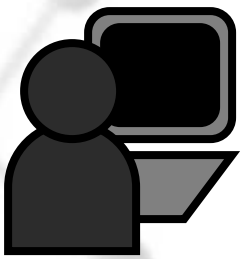
Multi-ID Users

letters

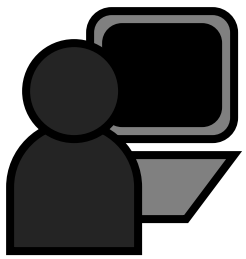
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
```

numbers

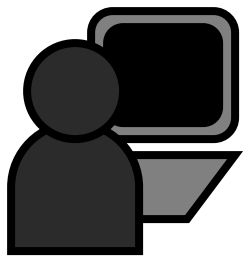
```
mack: i love 5

catt: i don't
joop: 27 rules!
```



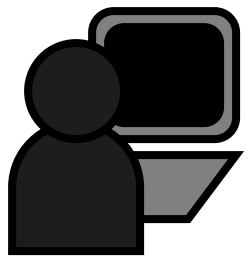
anna



mack



dogg/catt



joop

Rensselaer

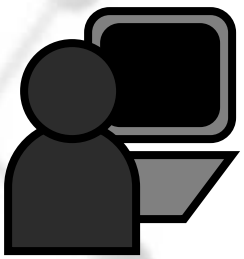
Multi-ID Users

letters

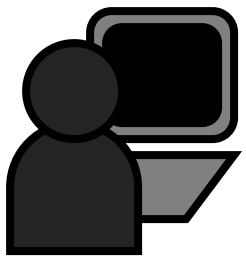
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
dogg: Z is the best
```

numbers

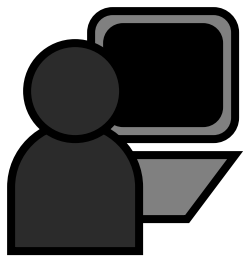
```
mack: i love 5

catt: i don't
joop: 27 rules!
```



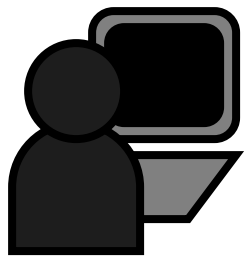
anna



mack



dogg/catt



joop

Rensselaer

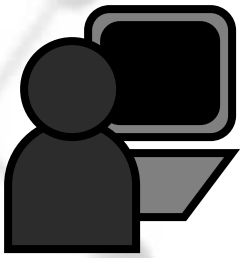
Multi-ID Users

letters

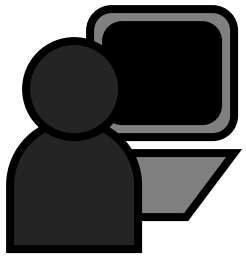
```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
dogg: Z is the best
```

numbers

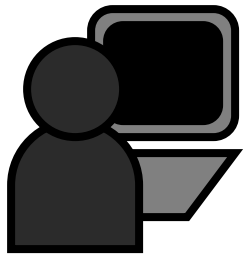
```
mack: i love 5

catt: i don't
joop: 27 rules!
catt: i agree joop :)
```



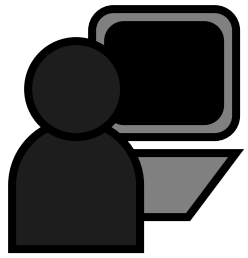
anna



mack



dogg/catt



joop

Rensselaer

Overview

- A model of a public forum
- Two efficient statistics-based algorithms for identification of multi-ID users
- Simulation results

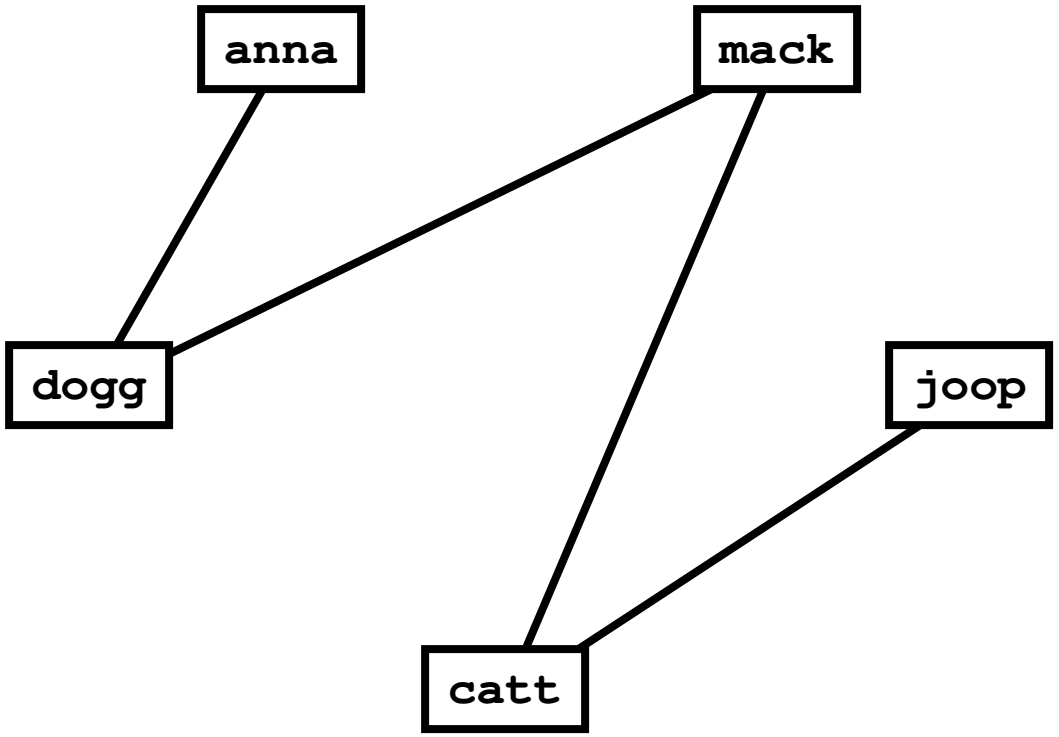
Overview

- A model of a public forum
- Two efficient statistics-based algorithms for identification of multi-ID users
- Simulation results

A Model of a Public Forum

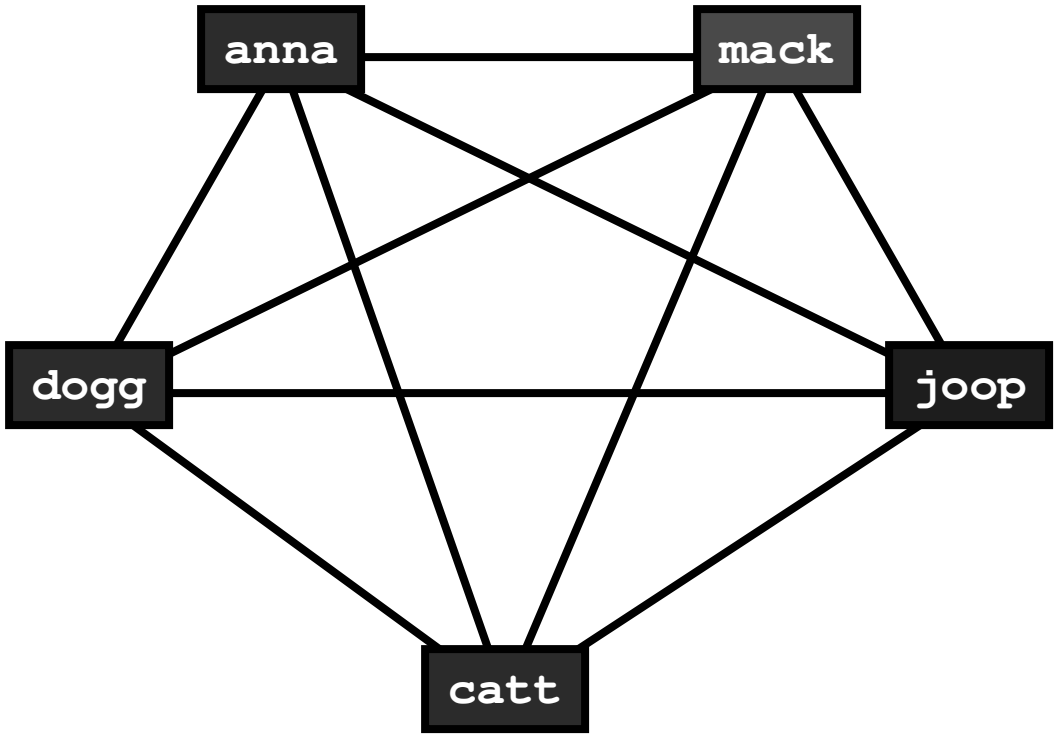
- Every actor has one response queue
- Common average and variance of response delay
- The server has a queue to process messages with very short delay

Model Friendship Graph



Rensselaer

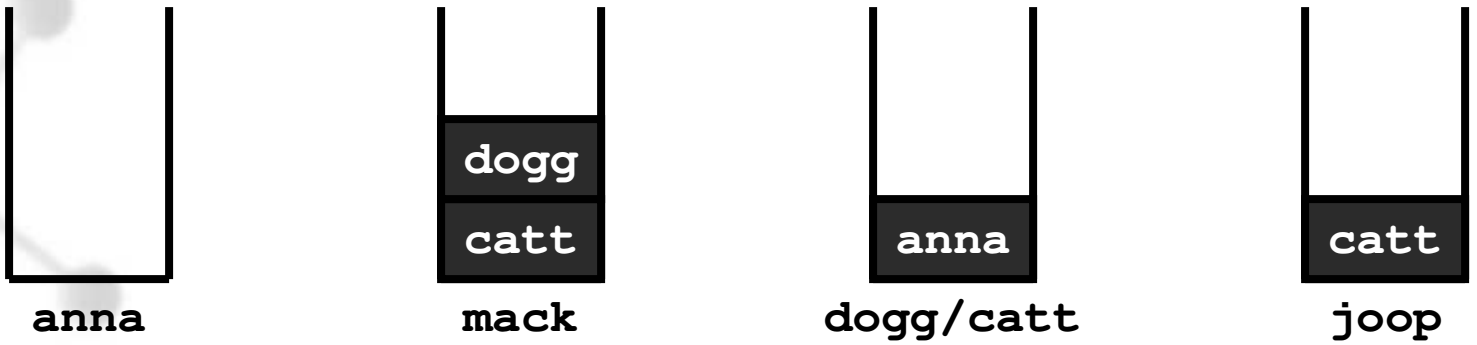
Model Alias Graph



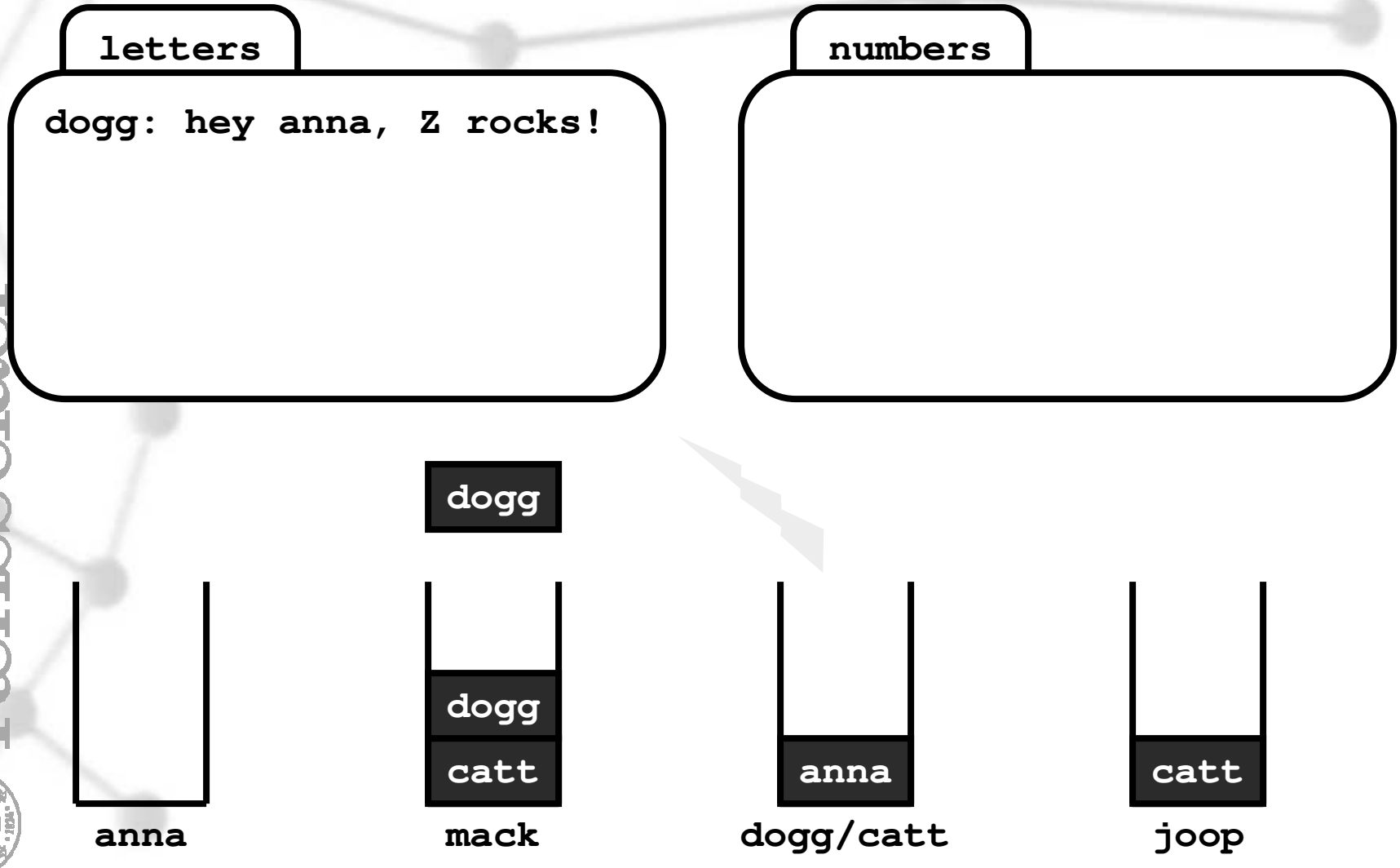
Rensselaer

Multi-ID Users

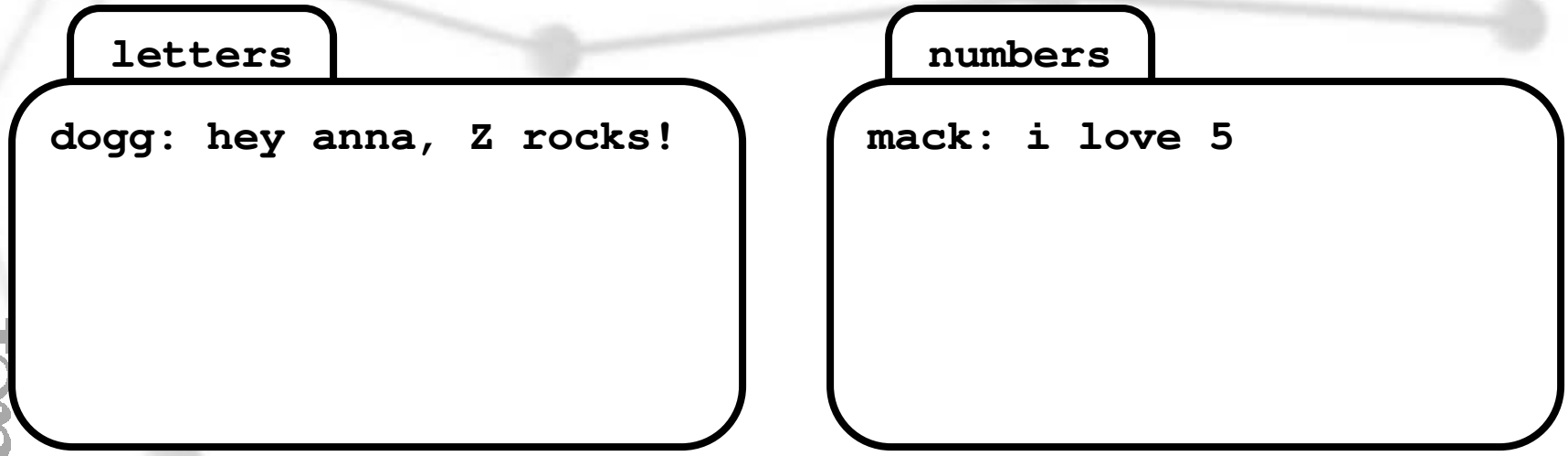
- One response queue for each actor

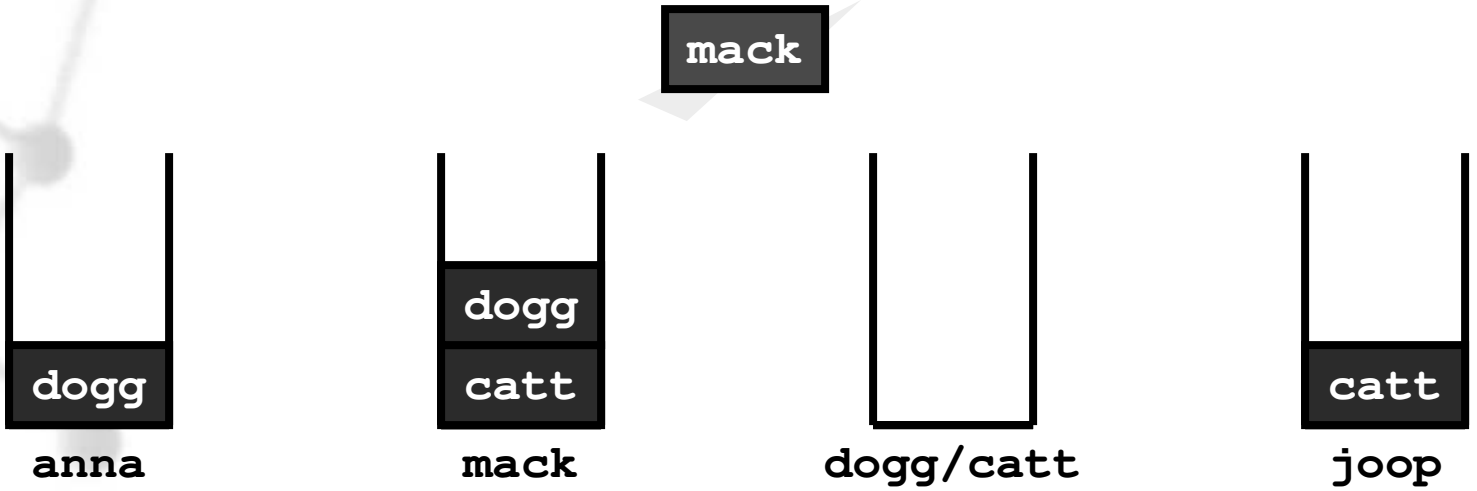


Multi-ID Users



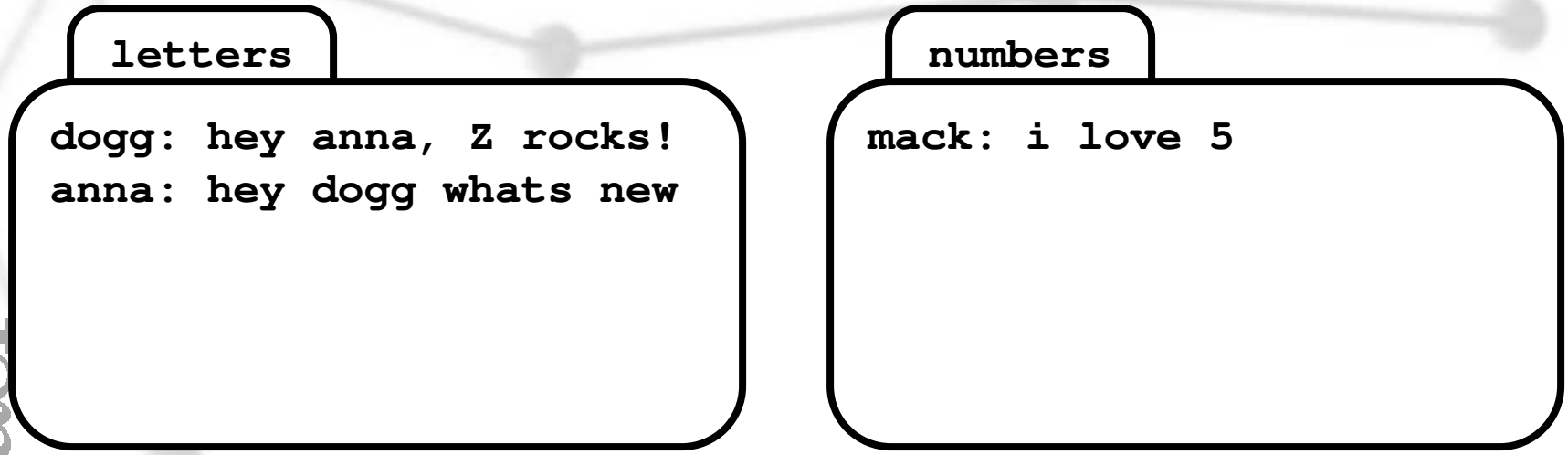
Multi-ID Users

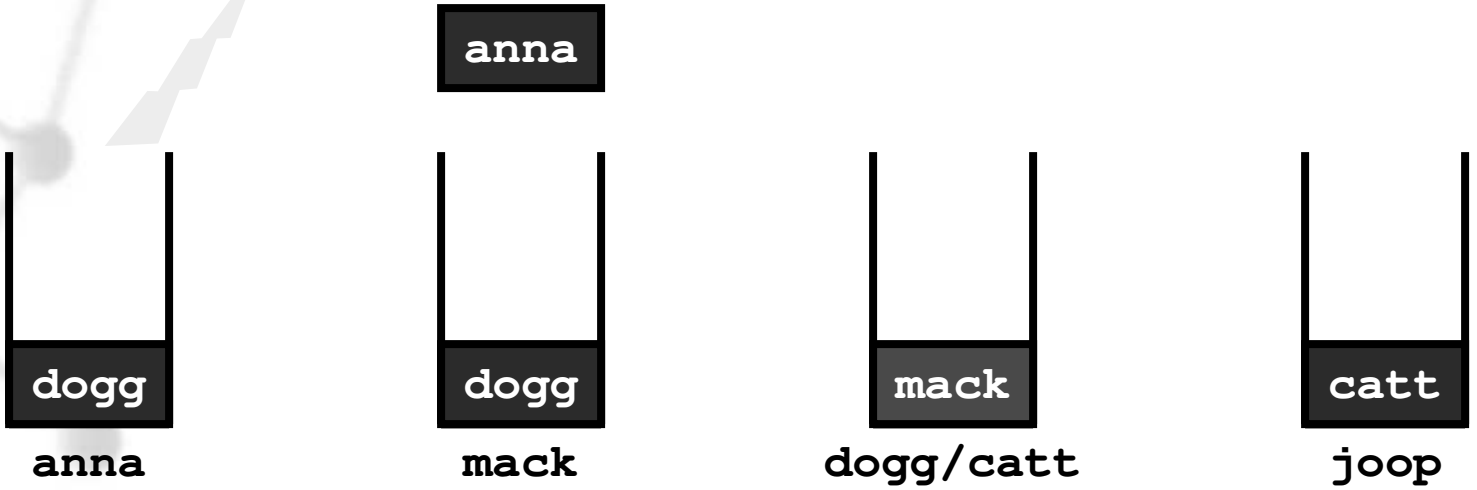




Rensselaer

Multi-ID Users





Rensselaer

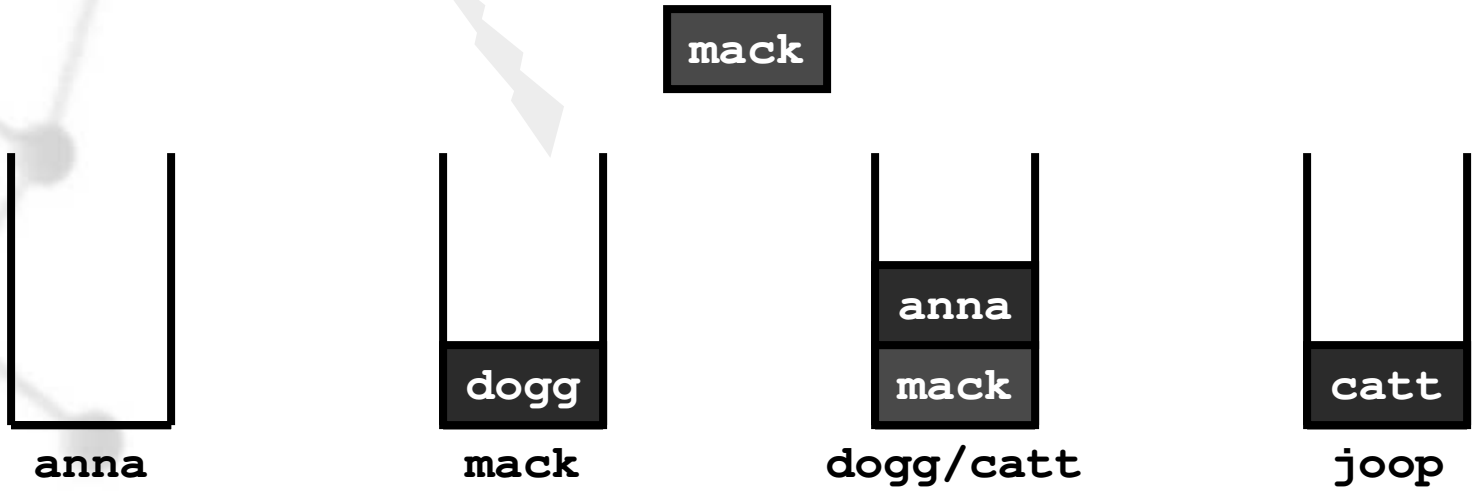
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
```

numbers

```
mack: i love 5
```



Rensselaer

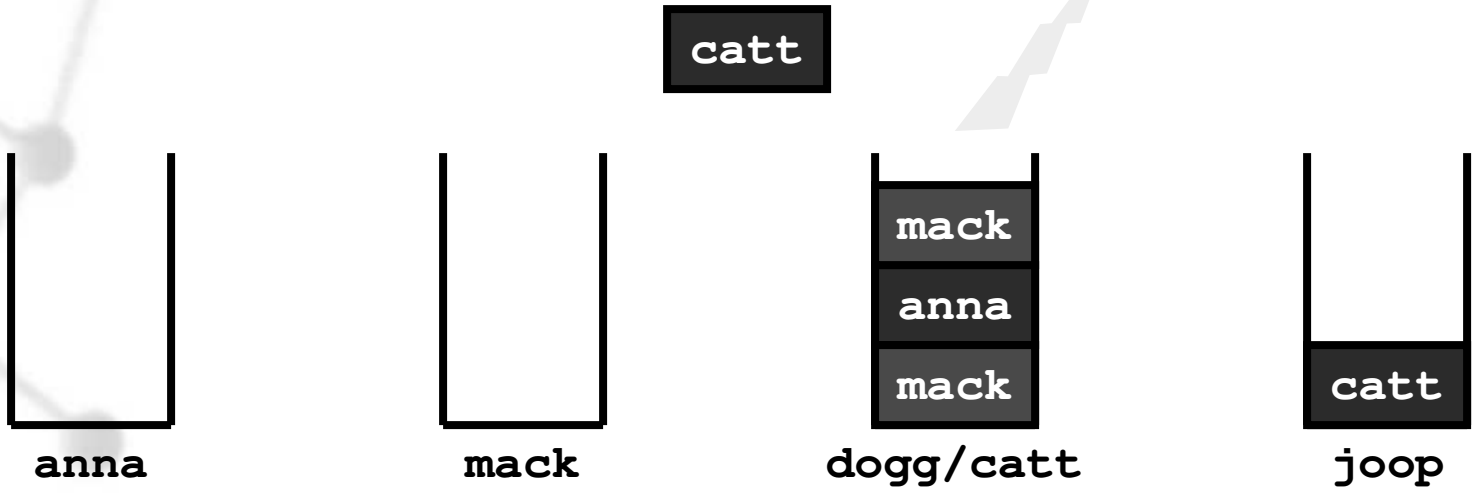
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
```

numbers

```
mack: i love 5

catt: i don't
```



Rensselaer

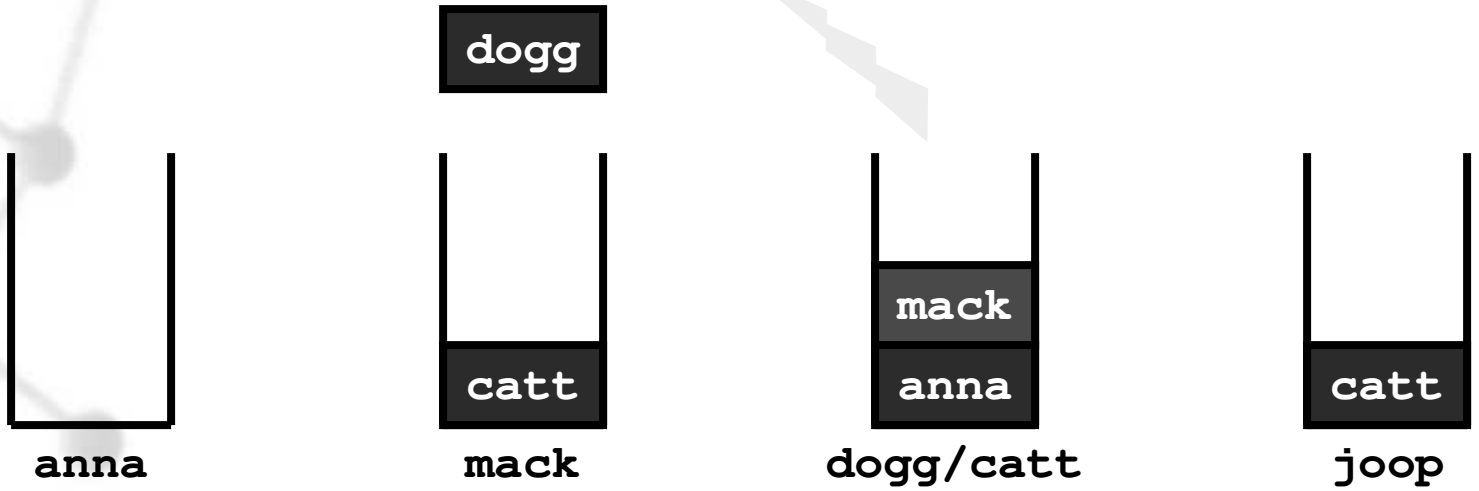
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
```

numbers

```
mack: i love 5

catt: i don't
```



Rensselaer

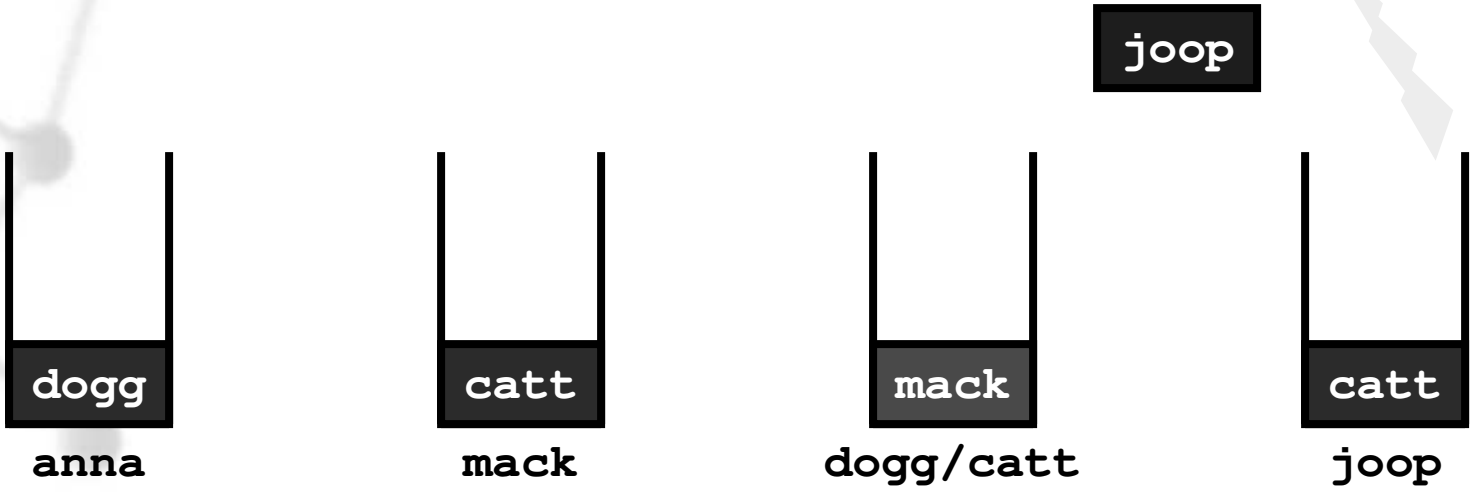
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
```

numbers

```
mack: i love 5

catt: i don't
joop: 27 rules!
```



Rensselaer

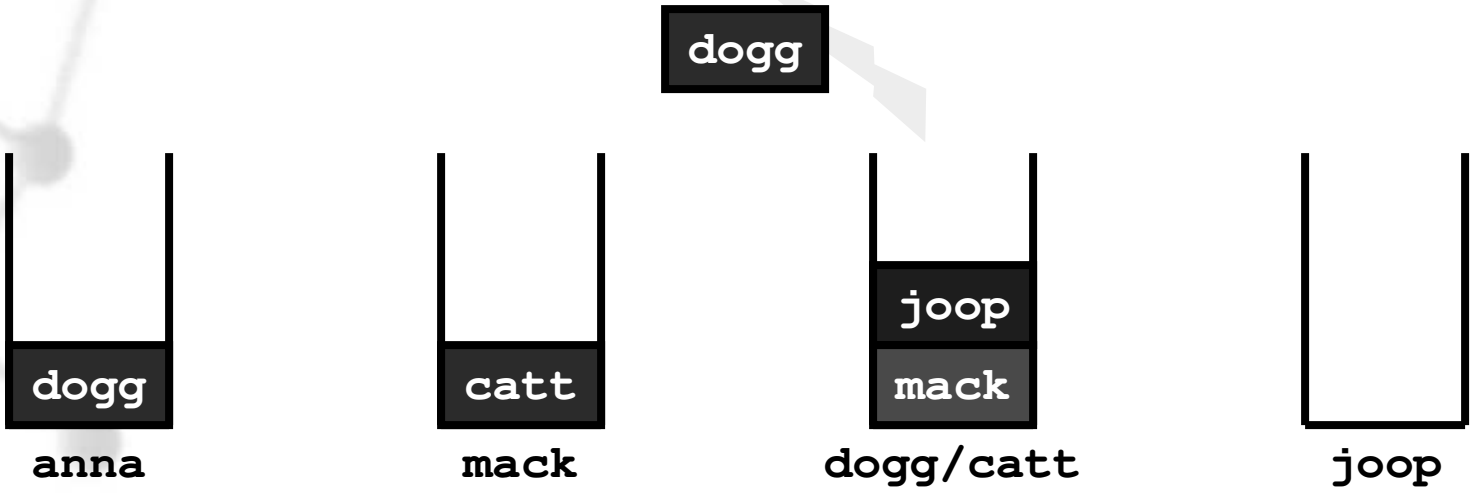
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
dogg: Z is the best
```

numbers

```
mack: i love 5

catt: i don't
joop: 27 rules!
```



Rensselaer

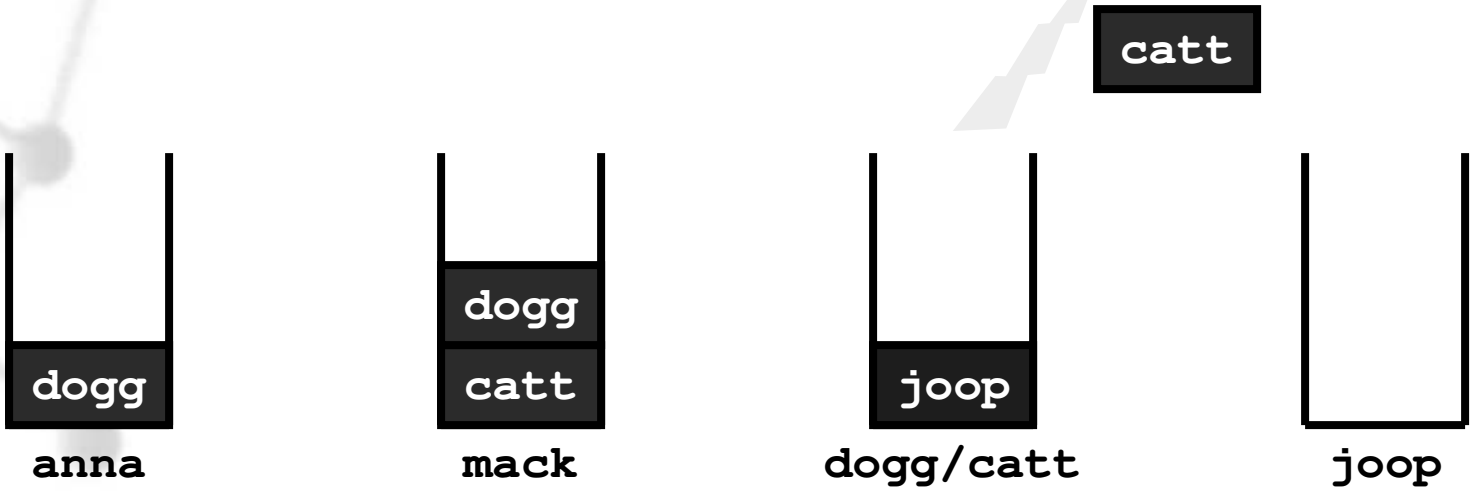
Multi-ID Users

letters

```
dogg: hey anna, Z rocks!
anna: hey dogg whats new
mack: what's that dogg?
dogg: not much
dogg: Z is the best
```

numbers

```
mack: i love 5

catt: i don't
joop: 27 rules!
catt: i agree joop :)
```



Rensselaer

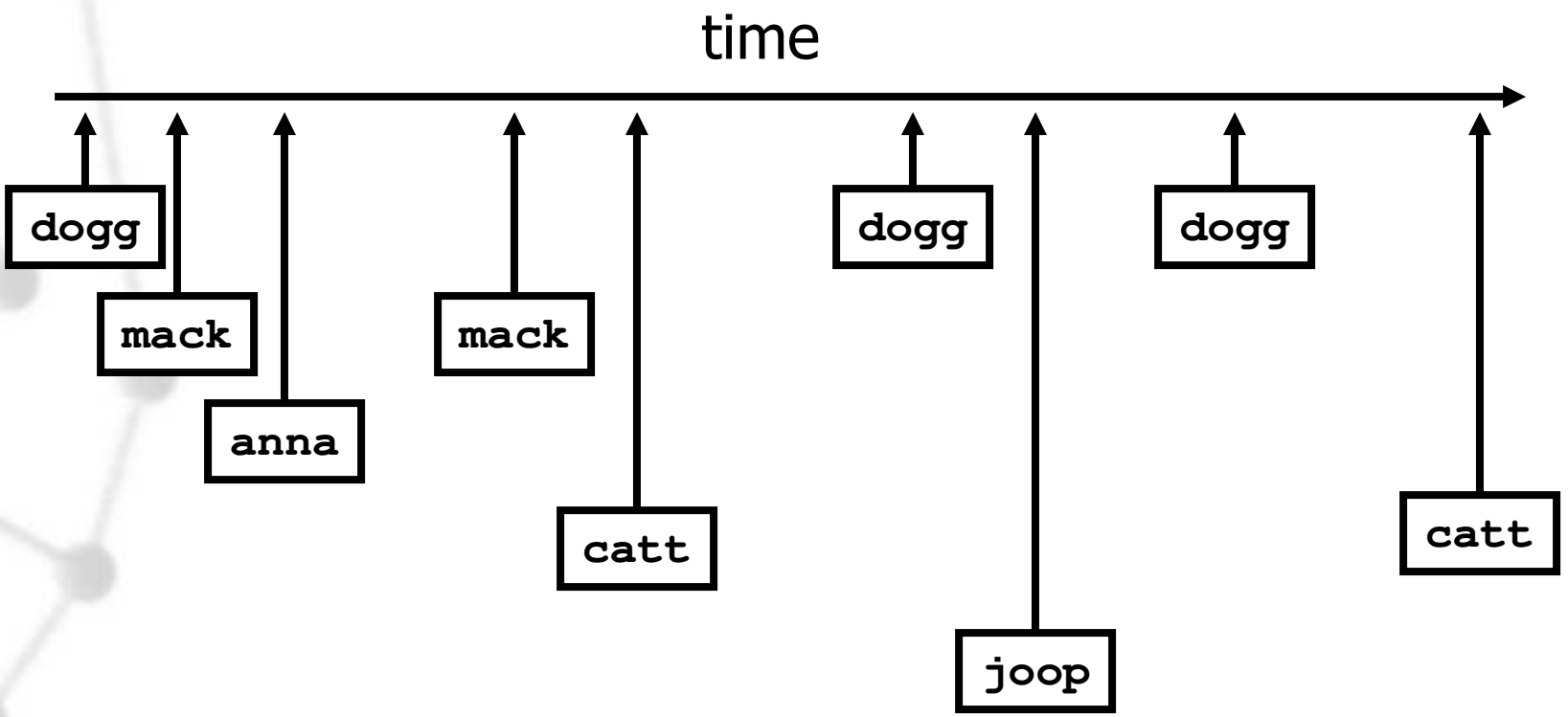
Overview

- A model of a public forum
- Two efficient statistics-based algorithms for identification of multi-ID users
- Simulation results

First Algorithm

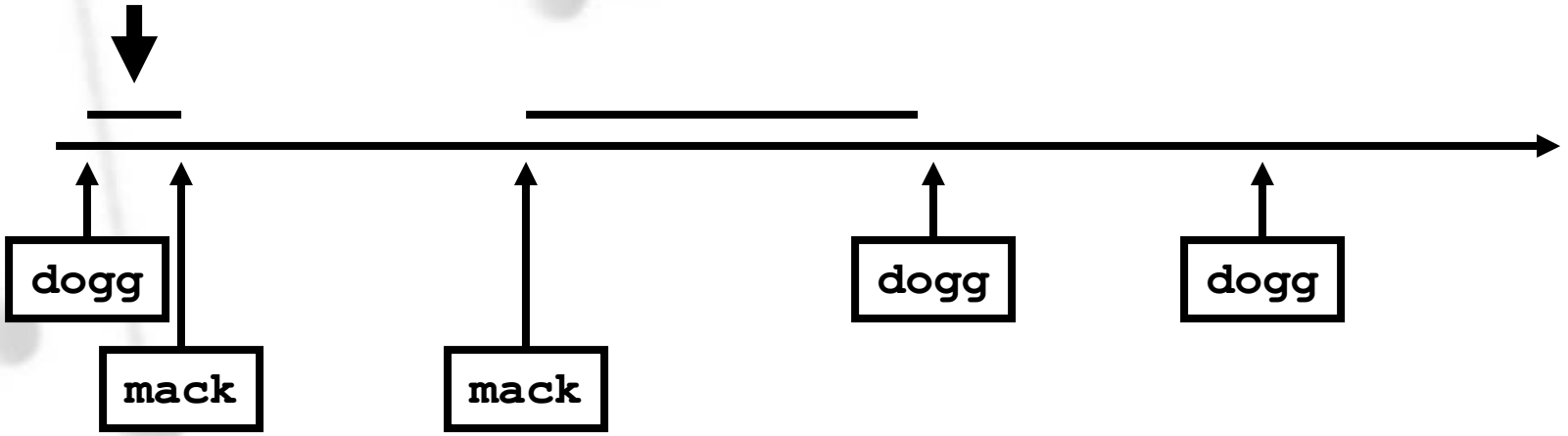
1. Collect information $\{\langle time_i, ID_i \rangle\}$
2. Compute minimum time delay minD for every pair of IDs
3. Cluster the delays using k-means into two groups.
   - Call pairs with larger center suspected to be the same actor (red)
   - Call pairs with smaller center suspected to different (blue)
4. Connected components using red edges are ID groups representing one actor

1. Collect information



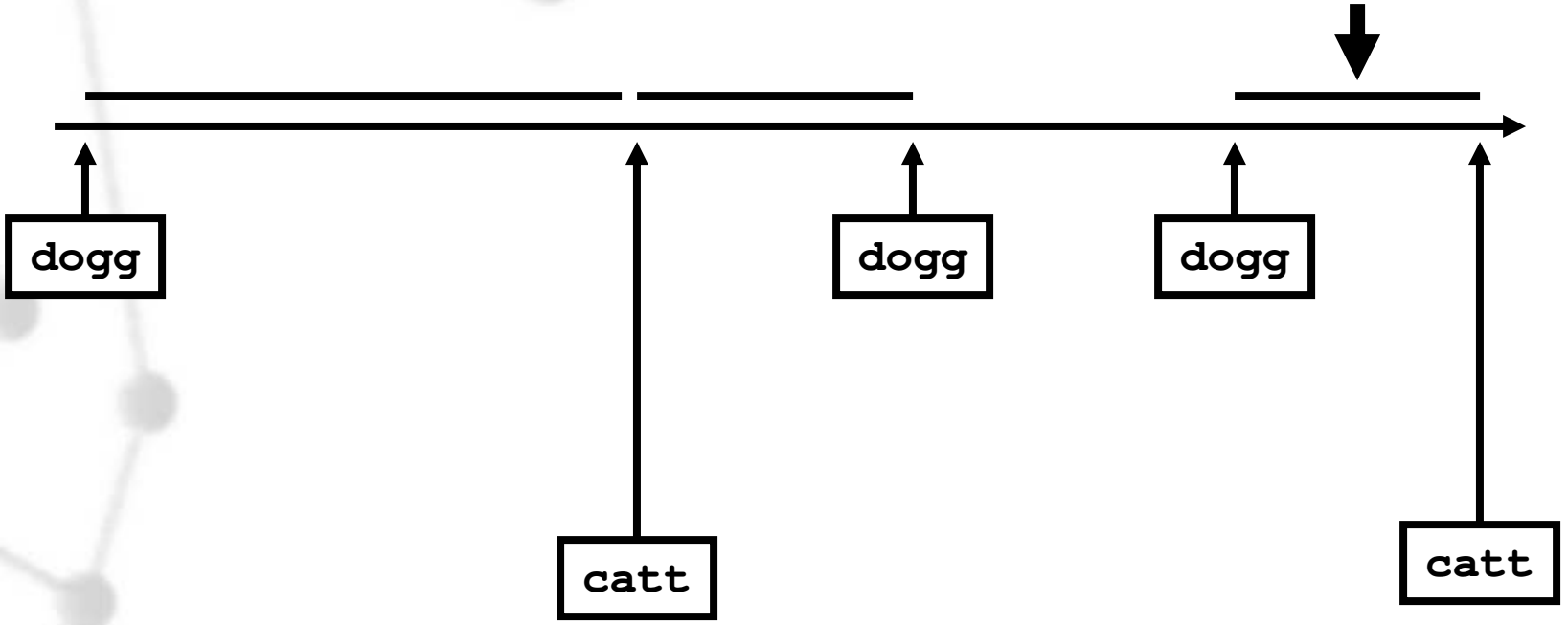
Rensselaer

2. minD(dogg, mack)



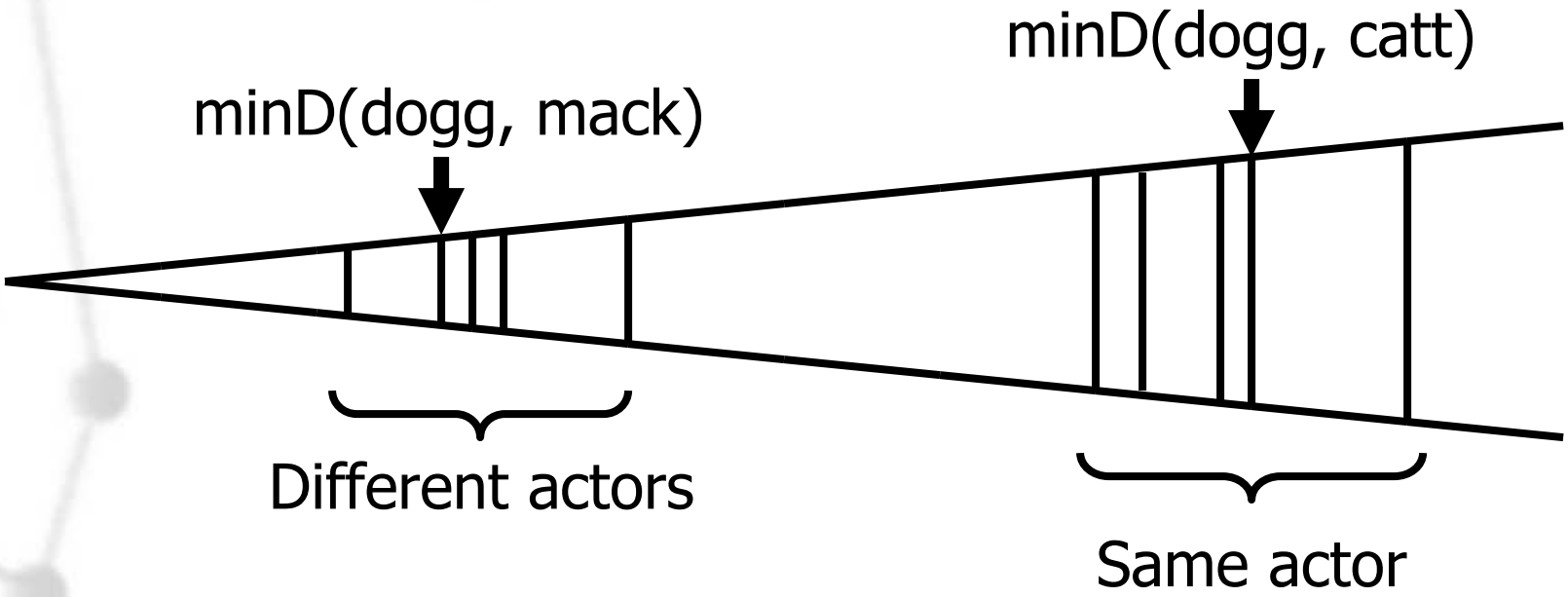
Rensselaer

2. minD(dogg, catt)



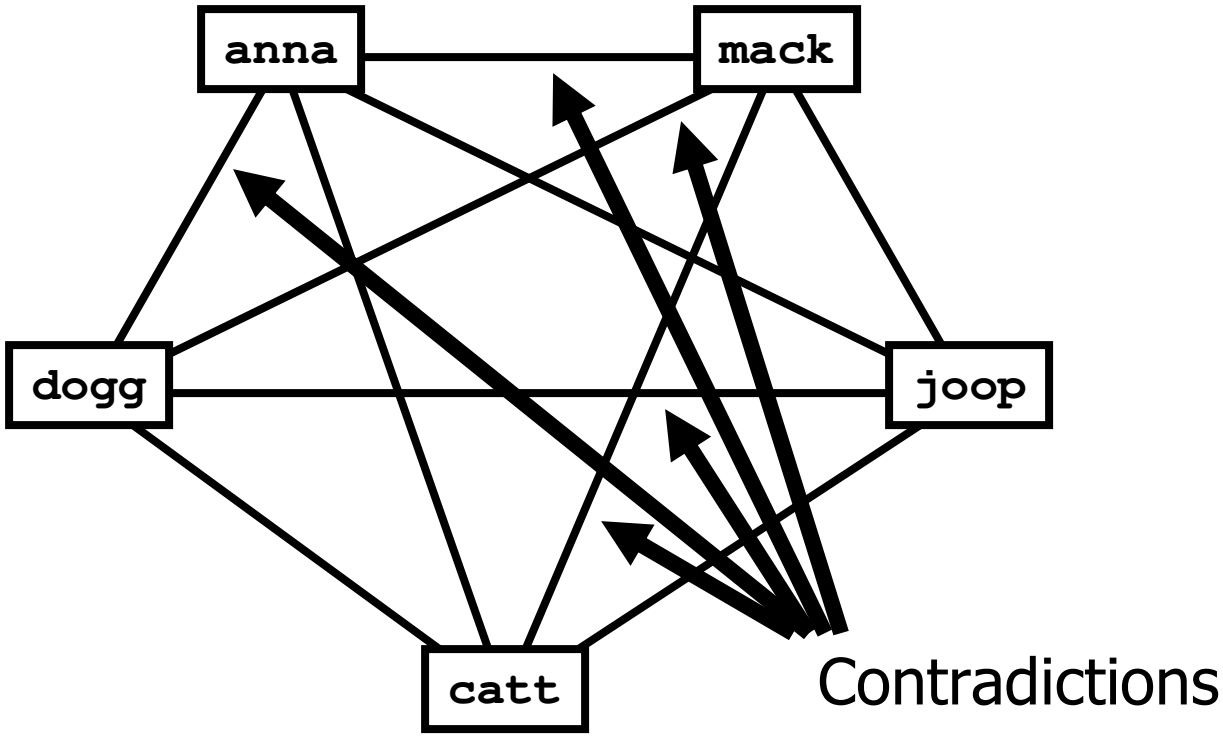
Rensselaer

3. Cluster minD values



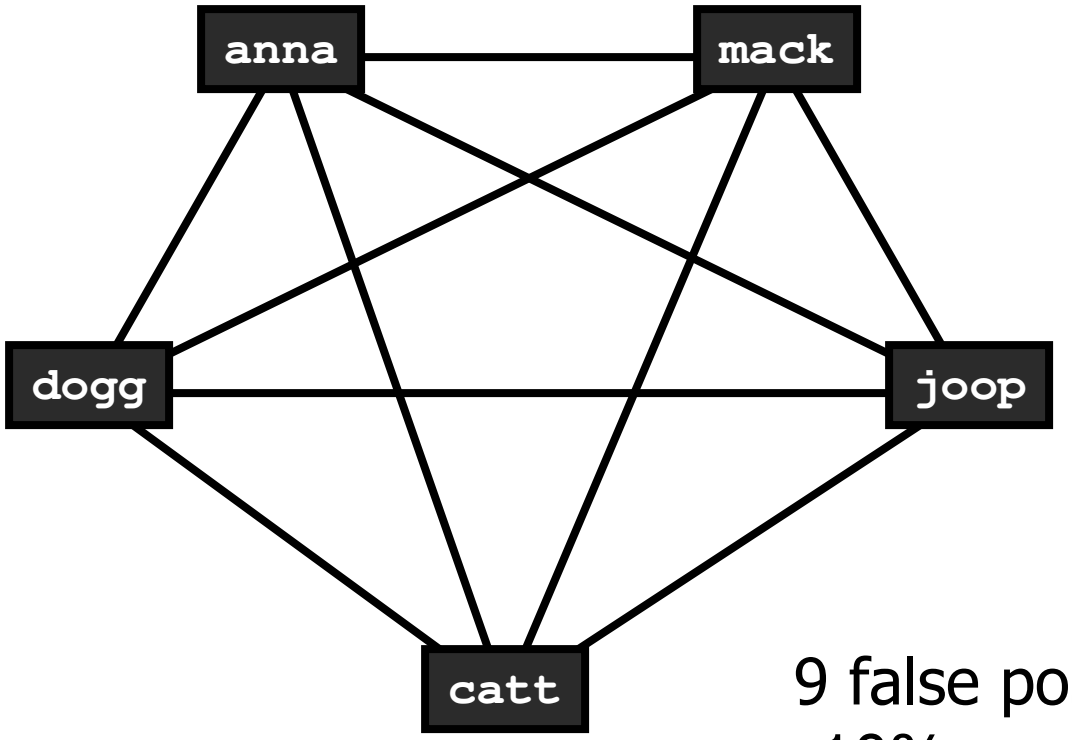
Rensselaer

3. Resulting Alias Graph



Rensselaer

4. One Red Component

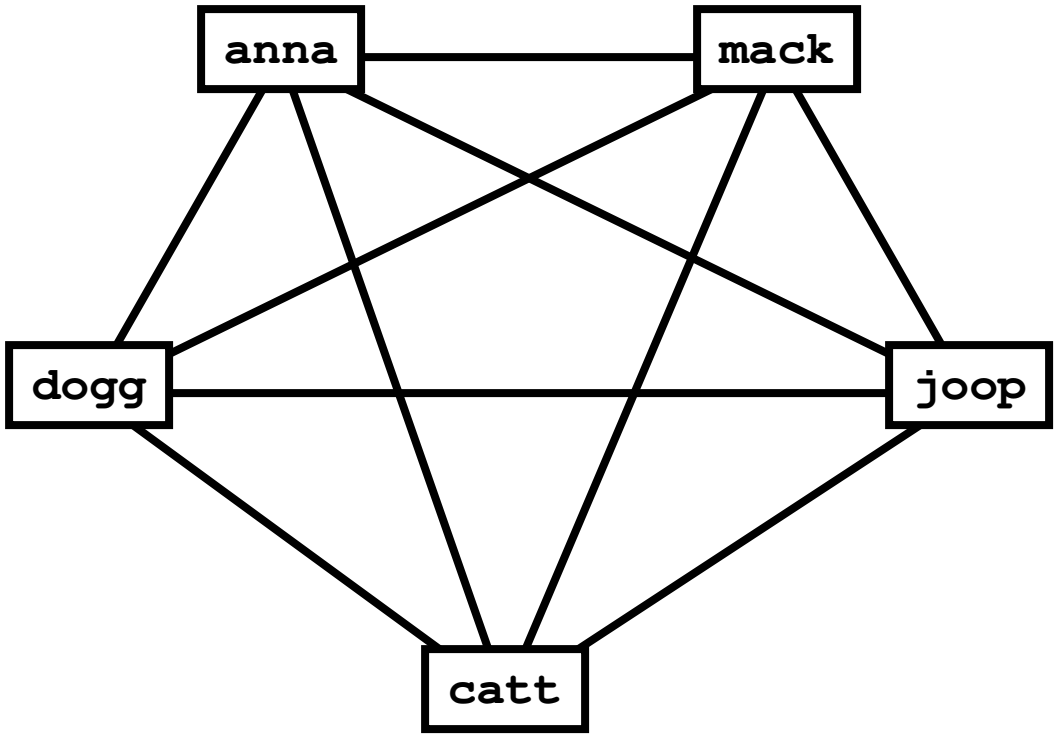


9 false positives
10% accuracy

Refinement Algorithm

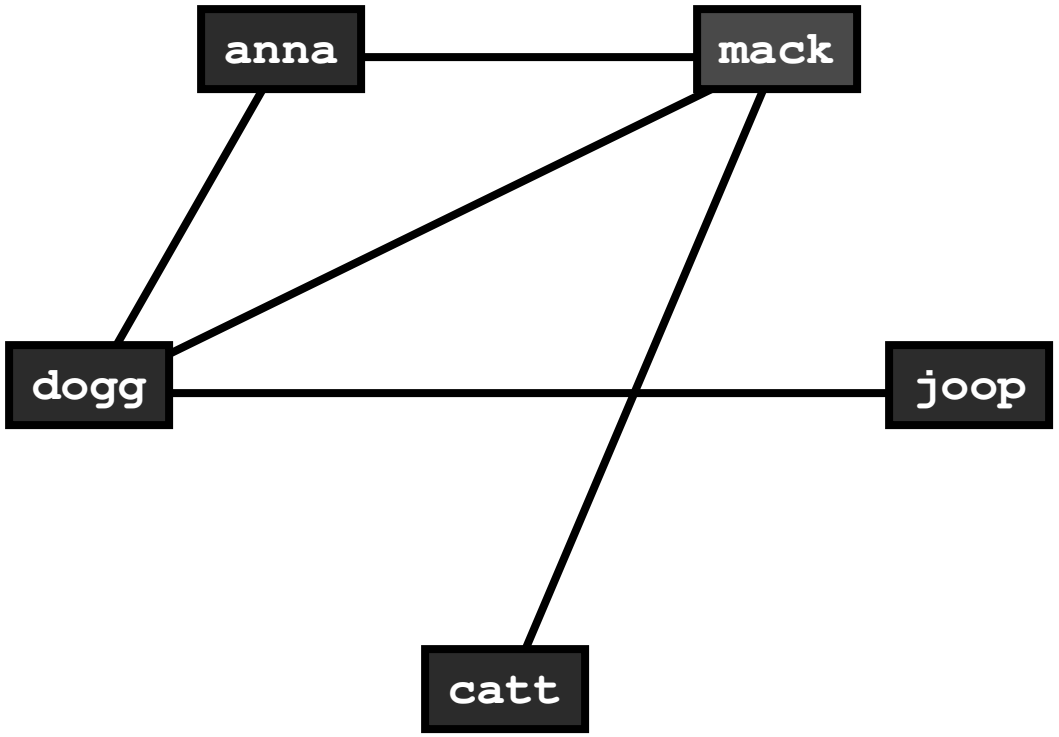
1.
2. } Follow original algorithm
3.
4. Find groups connected with red edges
5. Color IDs into smaller ID groups using blue edges

4. One Red Component



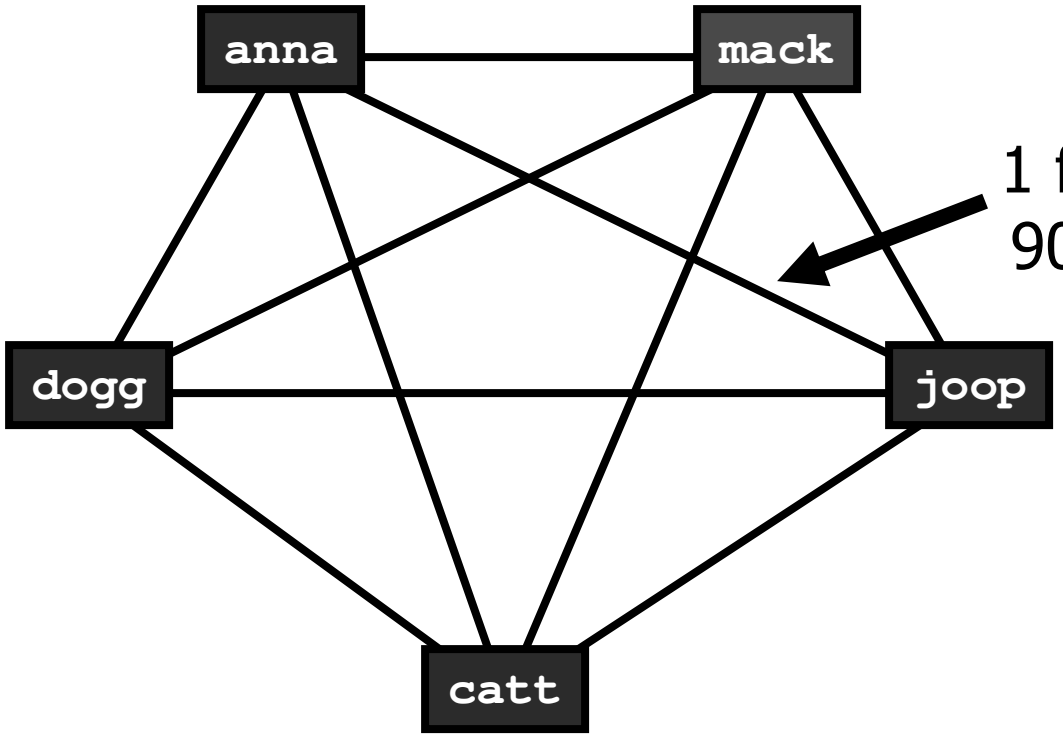
Rensselaer

5. Color Blue Graph



Rensselaer

5. Color Blue Graph



1 false positive
90% accuracy

Rensselaer

Overview

- A model of a public forum
- Two efficient statistics-based algorithms for identification of multi-ID users
- Simulation results

Simulation results

The average number of messages per pair	Accuracy without coloring (%)	Accuracy with coloring (%)
0.513	70.2797	70.9765
8.274	97.3140	97.3473
87.68	99.7762	99.7826
880.4	99.9995	99.9998

The dependence of Accuracy on the number of messages

Rensselaer

Future Work

- Real-life testing
- More sophisticated model
  - Different types of users
  - Short and overlapping online appearances
  - Length of posts
- Algorithm improvements
  - Statistical evaluation of the new models
  - Lexical and semantic analysis