Identifying Multi-ID Users in Open Forums

Hung-Ching Chen
Mark Goldberg
Malik Magdon-Ismail
Who is Using Multiple IDs?

Consider two chatrooms, “letters” and “numbers”

**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 :)
Multi-ID Users

dogg: hey anna, Z rocks!
Multi-ID Users

- **letters**
  - dogg: hey anna, Z rocks!

- **numbers**
  - mack: i love 5

- anna
- mack
- dogg/catt
- joop
Multi-ID Users

letters

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

numbers

mack: i love 5

anna
mack
dogg/catt
joop
Multi-ID Users

**letters**
- dogg: hey anna, Z rocks!
- anna: hey dogg what's new
- mack: what's that dogg?

**numbers**
- mack: i love 5
Multi-ID Users

letters

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

numbers

mack: i love 5
catt: i don't
Multi-ID Users

letters

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

numbers

mack: i love 5
catt: i don't
Multi-ID Users

letters

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

numbers

mack: i love 5
catt: i don't
joop: 27 rules!
Multi-ID Users

letters

dogg: hey anna, Z rocks!
anna: hey dogg what's 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!
Multi-ID Users

**Letters**

- dogg: hey anna, Z rocks!
- anna: hey dogg what's 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
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

- anna
- mack
- dogg
- catt
- joop
Model Alias Graph
Multi-ID Users

- One response queue for each actor
Multi-ID Users

letters

dogg: hey anna, Z rocks!

numbers

Rensselaer
Multi-ID Users

dogg: hey anna, Z rocks!

mack: i love 5
Multi-ID Users

letters

dogg: hey anna, Z rocks!
anna: hey dogg what's new

numbers

mack: i love 5
Multi-ID Users

letters

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

numbers

mack: i love 5
Multi-ID Users

letters

Dogg: hey anna, Z rocks!
Anna: hey dogg what's new
Mack: what's that dogg?

numbers

Mack: i love 5
Catt: i don't
Multi-ID Users

**letters**

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

**numbers**
mack: i love 5
catt: i don’t
Multi-ID Users

letters

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

numbers

- mack: i love 5
- catt: i don't
- joop: 27 rules!
Multi-ID Users

letters

dogg: hey anna, Z rocks!
anna: hey dogg what's 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!
Multi-ID Users

letters

dogg: hey anna, Z rocks!
anna: hey dogg what's 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 :)
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 \text{time}_i, \text{ID}_i \rangle\}
2. Compute minimum time delay \text{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 be different (blue)
4. Connected components using red edges are ID groups representing one actor
1. Collect information
2. minD(dogg, mack)
2. \text{minD(dogg, catt)}
3. Cluster minD values

- minD(dogg, mack)
  - Different actors

- minD(dogg, catt)
  - Same actor
3. Resulting Alias Graph

Contradictions

Diagram showing connections between nodes labeled "anna", "mack", "dogg", and "joop". Arrows indicate relationships or interactions between the nodes.
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
5. Color Blue Graph

![Graph diagram]

- Anna
- Mack
- Dogg
- Catt
- Joop
5. Color Blue Graph

1 false positive
90% accuracy
Overview

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

<table>
<thead>
<tr>
<th>The average number of messages per pair</th>
<th>Accuracy without coloring (%)</th>
<th>Accuracy with coloring (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.513</td>
<td>70.2797</td>
<td>70.9765</td>
</tr>
<tr>
<td>8.274</td>
<td>97.3140</td>
<td>97.3473</td>
</tr>
<tr>
<td>87.68</td>
<td>99.7762</td>
<td>99.7826</td>
</tr>
<tr>
<td>880.4</td>
<td>99.9995</td>
<td>99.9998</td>
</tr>
</tbody>
</table>

The dependence of Accuracy on the number of messages
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