

# Making Inferences from Individual Behavior Using Mobile Data



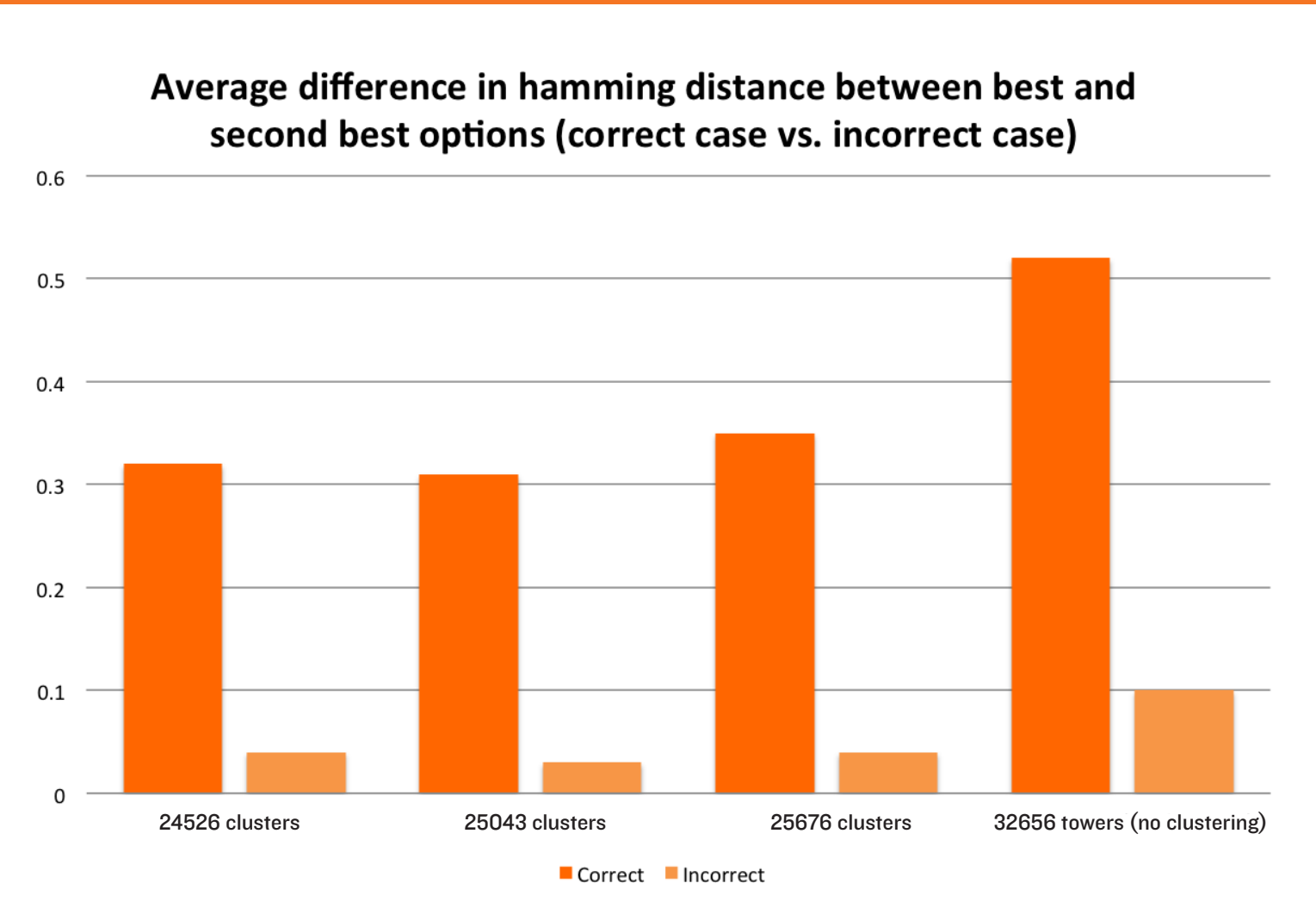
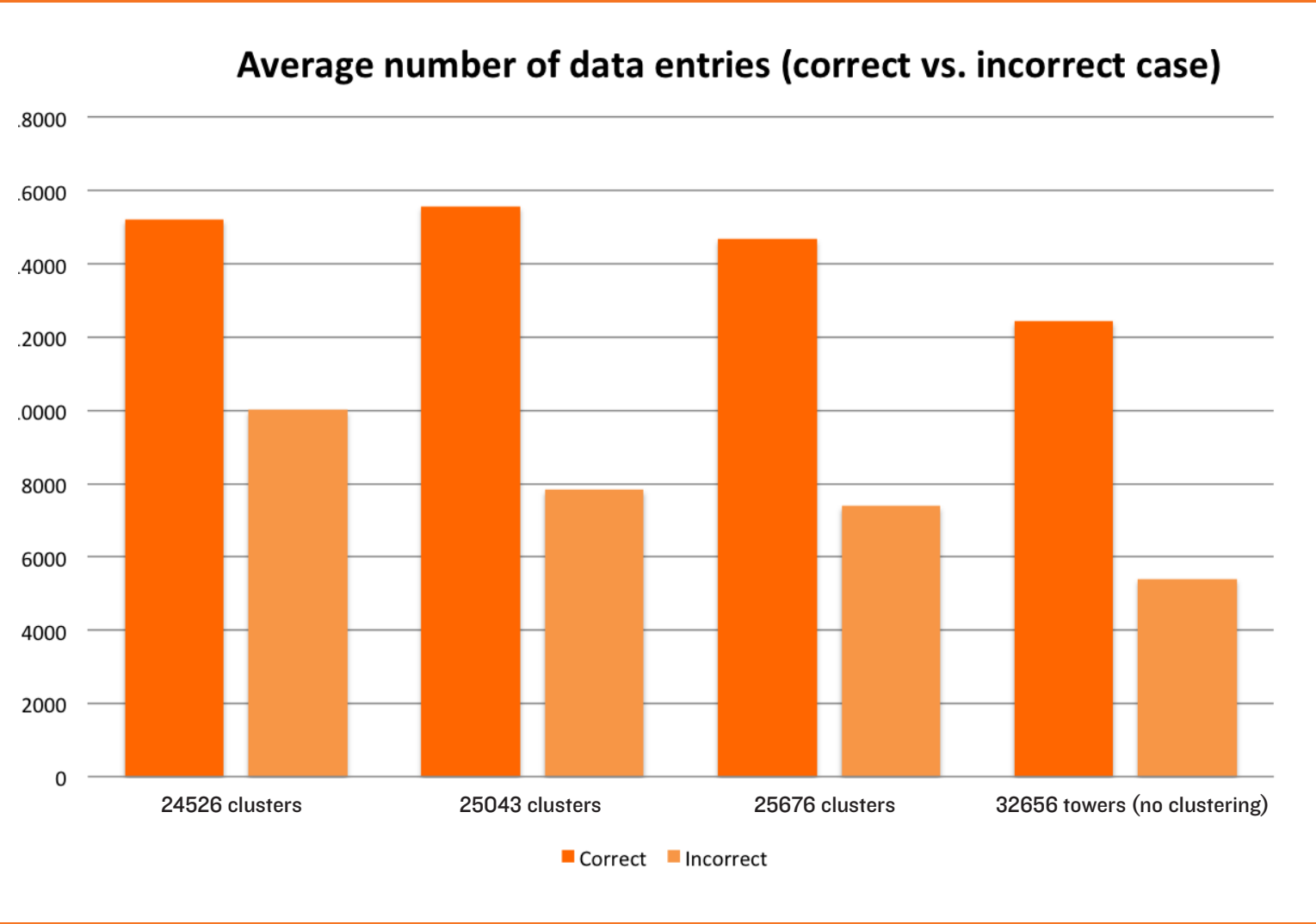
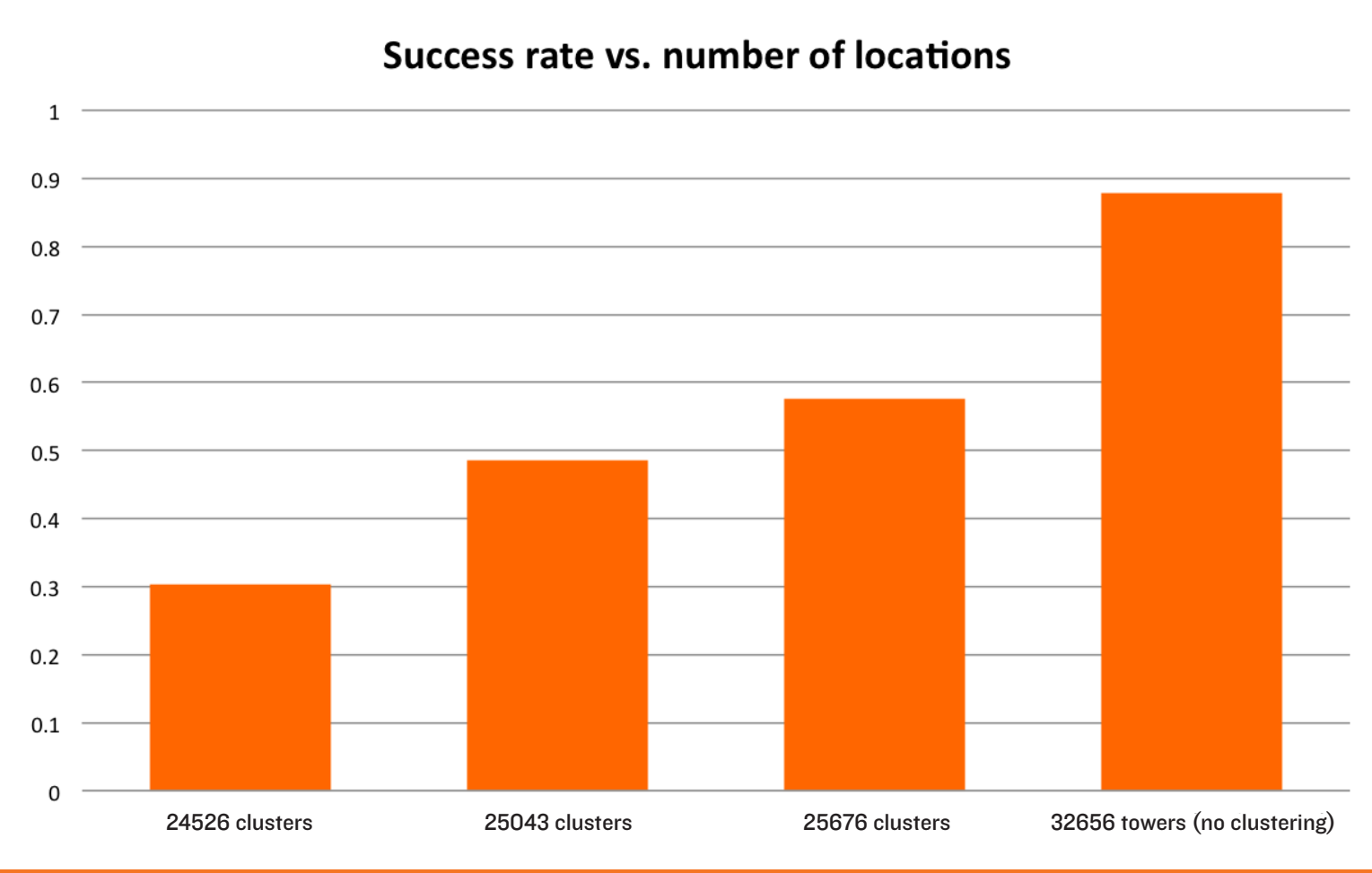
Tiantian Zha  
Arvind Narayanan

## De-anonymizing data

## Similarity between users

## Friendships

### Location data



### Goal

Given existing information about a set of users (e.g. Fall semester cell phone usage patterns) as well as anonymous data for one user from a different time period (e.g. usage patterns from the following Spring semester), **can we reconcile this anonymous user with an existing user?** We attempt to answer this question using 2 data sets: location data and voice calls / text messages data.

**Location:** We compute each user's distribution across the different locations, and compare the anonymous user's patterns with the known users' pattern. The metric used for comparison is the hamming distance. We iterate using different definitions of "location." Location can be defined as each cell tower, or as clusters of cell towers. We select the user with the lowest hamming distance from the anonymous user. **Phone:** We also find the similarity between the anonymous user and others.

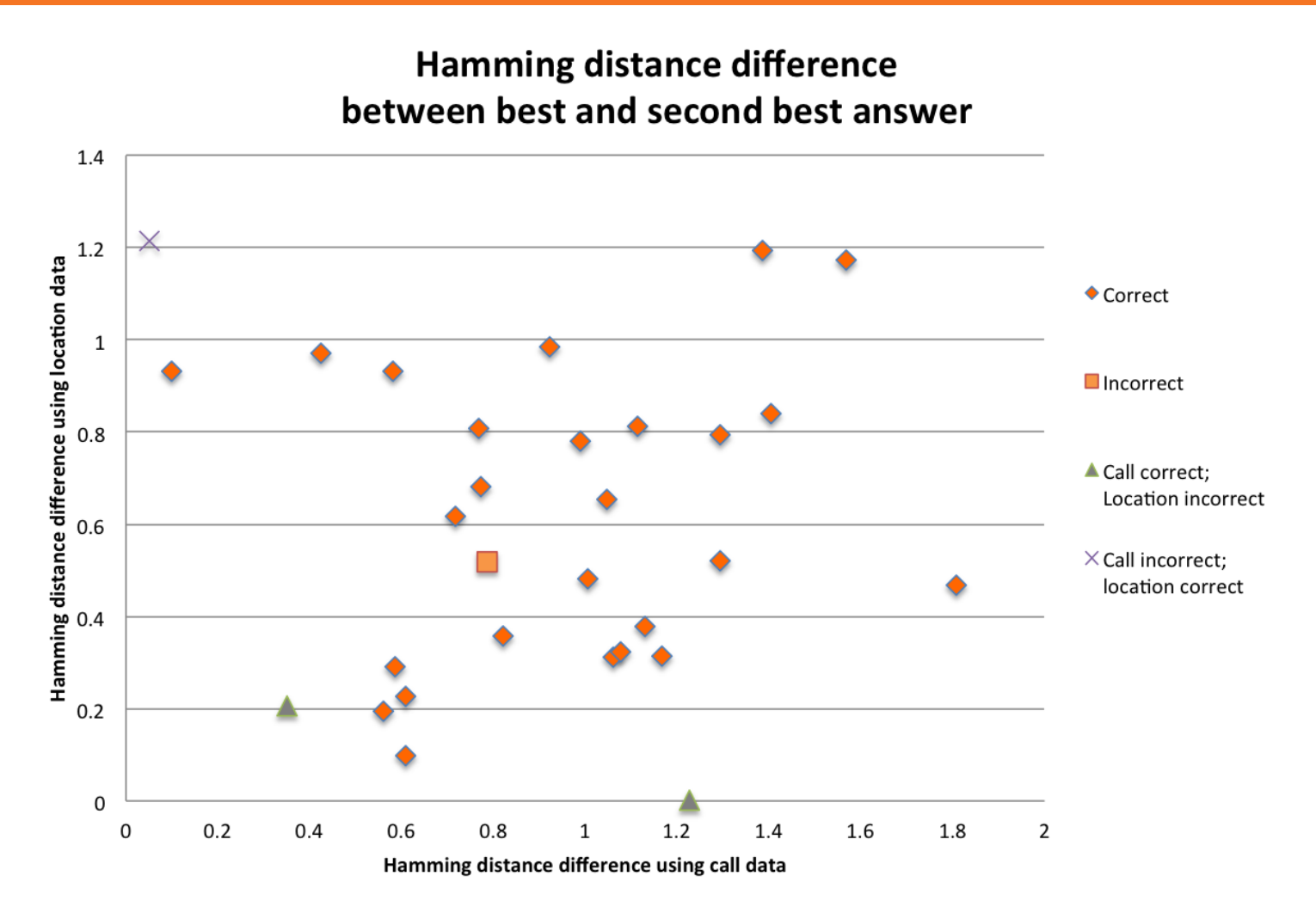
**Accuracy check:** We compare the hamming distance between the anonymous user and the first best choice with the hamming distance between the anonymous user and the second best choice.

### Call data

| Results summary table: success rate and amount of data for |       |                      |                        |
|--|-------|----------------------|------------------------|
| Type   | Rate  | Correct: amt of data | Incorrect: amt of data |
| Voice Call   | 28/30 | 4386 calls           | 1938 calls             |
| Short Message  | 15/25 | 972 msgs             | 335 msgs               |

| Avg. diff. in hamming distance between best guess and second best guess |  | Correct | Incorrect |
|---|--|---------|-----------|
|   |  | 0.96    | 0.42      |

### Accuracy check



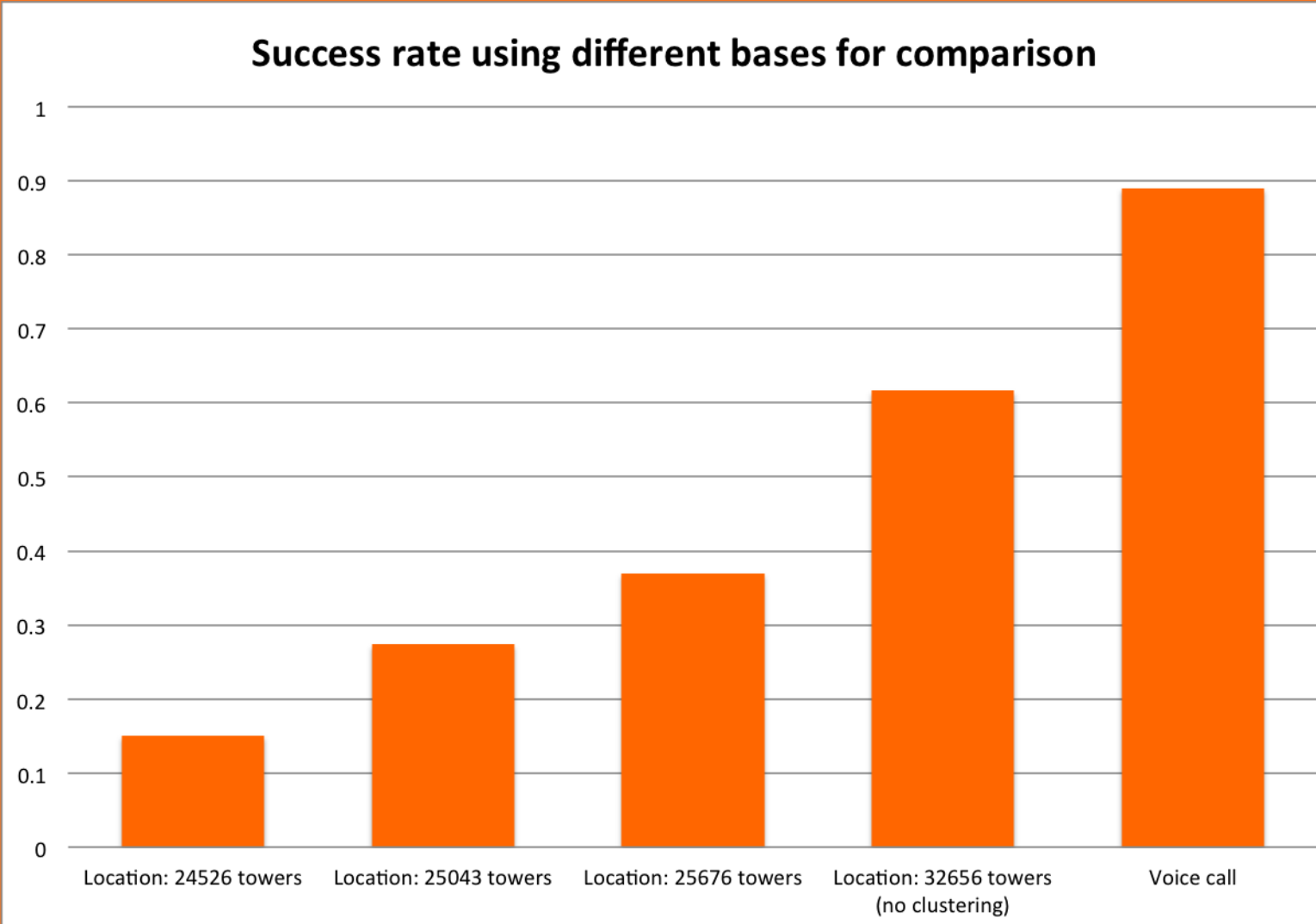
### Goal

The subjects of this study belong to different programs. They are **Media Lab graduate students, Media Lab undergraduate students, Sloan MBA students, and Media Lab staff/professors.**

Given one anonymous user, can we determine which program he/she belongs to, by comparing their similarity with other users? This question has privacy implications because we can infer information about this individual that is not implicitly disclosed.

We build "profiles" for each of the 4 groups, and compare the similarity between the anonymous user and each of the profiles. The metric for similarity was again the hamming distance between users' distributions.

### Results



### Analysis

We were much more successful in making predictions about the anonymous user when using their voice call data than when using location data. To understand why, we computed the differences between the 4 profiles. The profiles built using location data tended to resemble each other more than the profiles built using voice call data. This makes sense, as we expect many students to share the same facilities, whereas we expect that different groups will have different social circles.

| Diff. between profiles built using location (left) and calls (right) |         |              |       |          |
|--|---------|--------------|-------|----------|
| Profiles   | ML grad | ML undergrad | Sloan | ML staff |
| ML grad  | -       | 1.216        | 1.262 | 1.435    |
| ML undergrad   | 1.216   | -            | 1.432 | 1.540    |
| Sloan  | 1.262   | 1.432        | -     | 1.654    |
| ML staff   | 1.435   | 1.540        | 1.654 | -        |

| Profiles     | ML grad | ML undergrad | Sloan | ML staff |
|--------------|---------|--------------|-------|----------|
| ML grad      | -       | 1.898        | 1.907 | 1.894    |
| ML undergrad | 1.898   | -            | 1.913 | 1.868    |
| Sloan        | 1.907   | 1.913        | -     | 1.888    |
| ML staff     | 1.894   | 1.868        | 1.888 | -        |

### Goal & Results

We hoped to determine whether it was possible to infer friendships from time spent together.

We calculated the amount of time users spent in each others' presence, and ranked, for each user, all other users in descending order of time spent together. **We hypothesized that the higher the rank, the more likely they were to be friends.**

We tested this hypothesis by using the 62 reported pairs of friendships, and generating at random 62 other pairs. If our algorithm could accurately find the 62 real friend pairs, then the amount of time spent together was an effective metric.

**Results.** We were able to locate 33 of the 62 true pairs of friends out of the entire pool. This result is about as effective as a random guess, since we expect random guessing to select 31 out of the 62 pairs. Therefore, amount of time spent together is a poor metric.

It is likely that there are **too many factors contributing to proximity**, other than friendship, and simply using the raw amount of time spent as a proxy is too crude of a metric. Other metrics to try would be to include the diversity of locations where two users are together, for example.

## Reality Mining Data

We performed the above analysis on the Reality Mining Database provided by the MIT's Media Lab. The database followed **97 subjects for 9 months** and tracked their cell phone usage.

This database contains:

- Calls and messaging
- Location by cell tower
- Semantic names given to each cell tower by a user
- Survey data from each user
- Other information to help link data together

Specifically, we selected data of events that took place during MIT's two academic semesters, from August 30, 2004 to December 17, 2004 (Fall semester) and from February 1 2005 to May 20 2005 (Spring semester). This allows us to make the users more easily comparable.