**Introduction**

Police often recover suspects’ phones during criminal investigations. Manually examining phone data to uncover relationships between individuals, however, is both difficult and time-consuming. In our project, we modeled and predicted relationships between individuals based only on their call and text logs. In particular, we focused on friend and couple relationships. Successful extensions of our work could be of great help to police in their investigations.

**Research Questions**

- What call and text information is useful for determining whether two individuals are friends?
- Can we predict whether two individuals are a couple?

**Materials and Methods**

**Data Collection**

- We used cell phone data collected by MIT’s Human Dynamics Lab.
- Data were collected from 200 participants in residential living community from March 2010 – September 2011.
- Data included information about call/text activity:
  - Number of calls/texts in/out between the individuals
  - Proportion of calls/texts in/out between the individuals
  - Number of people the first individual is in contact with
  - Total calls/texts made by the first individual

**Friend Pair Dataset:**

- Data were labeled by assigned friend score (0 is lowest, 7 is highest) or strength of friendship (0 is weak, 5.7 is strong).
- We used logistic regression and principal component analysis (PCA) to infer significant predictors of friendship strength.

**Couple Pair Dataset:**

- Data were labeled by whether the pair of individuals are a couple:
  - Only 36 couple relationships out of 6,000+ relationships
- We used logistic regression and PCA to predict couple status.

**Text and Call Relationships**

- Scatterplot of calls out vs. texts out in a given relationship.
- In general, calling someone more is associated with texting them more as well. Notably, though, a significant number of relationships do not involve any texts at all.

**Friend Score Relationships**

- Relationships with more calls in/out have higher friend scores.

**Friend Score Modeling**

We train logistic regression models to gain further insight into some of the more complex call/text behaviors associated with friendship.

**Model 1:** Strong Friendship

\[ \text{Is Popular} + \text{Proportion of Tests Out + } \text{(Is Popular + Proportion of Tests Out)} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Popular</td>
<td>0.197</td>
<td>0.007</td>
</tr>
<tr>
<td>Proportion of Tests Out</td>
<td>4.966</td>
<td>0.007</td>
</tr>
<tr>
<td>Is Popular * Proportion of Tests Out</td>
<td>72.054</td>
<td>0.003</td>
</tr>
</tbody>
</table>

For participants in contact with many others, having a high proportion of tests out to a specific person is a much stronger indicator of a strong friendship than it is for participants in contact with fewer others.

**Model 2 (PCA):**

\[ \text{Participant ID} + \text{PC1 + PC2 + PC3} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>0.470</td>
<td>0.00004</td>
</tr>
<tr>
<td>PC2</td>
<td>-4.20</td>
<td>0.001</td>
</tr>
<tr>
<td>PC3</td>
<td>1.68</td>
<td>0.007</td>
</tr>
</tbody>
</table>

The three principal components can be interpreted as follows:

1) **Volume of tests and calls between the pair of people**
2) **Each individual's propensity to text rather than call**
3) **Volume of texts, rather than calls, between the pair of people**

In general, calling and texting more is associated with an increased odds of being strong friends. While individuals who text a lot tend to have weaker relationships in general, texting someone a lot is indicative of a strong friendship.

**Couple Pair Relationships**

- Couples tend to have a higher proportion of calls in and out with each other.
- Proportion of total calls in and calls out can effectively separate couples from non-couples.

**Model 3:**

\[ \text{Is Couple ~ (Variables in table)} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-3.52</td>
<td>2.13</td>
</tr>
<tr>
<td>Calls In-Night</td>
<td>-2.12</td>
<td>0.28</td>
</tr>
<tr>
<td>Calls Out-Night</td>
<td>1.37</td>
<td>0.94</td>
</tr>
<tr>
<td>Calls In-Morning</td>
<td>-2.49</td>
<td>0.80</td>
</tr>
<tr>
<td>Calls Out-Morning</td>
<td>-3.69</td>
<td>0.94</td>
</tr>
<tr>
<td>SMS In-Night</td>
<td>2.99</td>
<td>0.94</td>
</tr>
<tr>
<td>SMS Out-Night</td>
<td>2.39</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Model 4 (PCA):**

\[ \text{Is Couple ~ (PC1 + PC2 + PC3)} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.89</td>
<td>0.82</td>
</tr>
<tr>
<td>PC1</td>
<td>0.28</td>
<td>0.82</td>
</tr>
<tr>
<td>PC2</td>
<td>0.49</td>
<td>0.82</td>
</tr>
<tr>
<td>PC3</td>
<td>0.21</td>
<td>0.82</td>
</tr>
</tbody>
</table>

**Conclusion**

- Overall volume of calls and texts in a relationship can account for a significant portion of the strength of the relationship.
  - Individuals in contact with many others tend to contact more medium-level friends.
  - People disproportionately test their closest friends more.
- Couple relationships are easier to identify than friendships.
  - A high proportion of calls in/out is a good indicator of a couple relationship.

If there are any further questions, feel free to contact Jonathan Che at jche818@amherst.edu

**Acknowledgements**

We would like to thank our TA, Ciaran Evans, for all of his assistance, as well as the CMU Statistics Department, CSAFE, and NIST.

---

**Discussion**

In our project, we used cell phone call and text logs to model relationships between different people. A number of general trends emerged; for example, friend relationships that involved more calls and texts tended to be reported as stronger relationships.

We also found that the individuals who are in contact with many people seem to have more medium-strength relationships. The strength of these “popular” individuals’ relationships is also more heavily related to their text activity.

Finally, we trained models to predict whether a given relationship is a couple relationship. Weighted logistic regression models performed decently well on the data, achieving high sensitivity while not predicting too many false positives. Interestingly, PCA helps separate couples from non-couples, which may indicate that couples just have significantly more contact with each other in some sense.

**Limitations/Next Steps**

- Dependencies due to network structure of data
- Unexplained errors/discrepancies in data collection process
- Network modeling to account for dependencies
- Community detection

**Figure 1:** Network graph of call and text data, node size proportional to degree. Some individuals are contacted much more than others.

**Figure 2:** Relationships with more calls in/out have higher friend scores.

**Figure 3:** Participants in contact with many people tend to assign friend scores in the 4-6 range, while participants in contact with fewer people tend to assign friend scores of either 1 or 7.

**Figure 4:** Couples tend to have a higher proportion of calls in and out with each other.

**Figure 5:** Network graph of strong friend relationships, colored by detected communities.