1 Definitions

1.1 Friendship:

Defined as having met another node at least twice after 5 PM or having met another node at least twice in a context that is not "university."

1.2 Personalized Page Rank

- Briefly, a personalized PageRank is like standard PageRank, except that when randomly teleporting to a new node, the surfer always teleports back to the given source node being personalized (rather than to a node chosen uniformly at random, as in the classic PageRank algorithm).
- That is, the random surfer in the personalized PageRank model works as follows:
- He starts at the source node X that we want to calculate a personalized PageRank around. At step i: with probability p, the surfer moves to a neighboring node chosen uniformly at random; with probability 1−p, the surfer instead teleports back to the original source node X. The limiting probability that the surfer is at node N is then the personalized PageRank score of node N around X.

1.3 Mean Average Precision

Some simplification of the Mean Average Precision measure is used to score the predictions. At the moment though order doesn't matter.

2 Data

I split the whole observation period into bins of roughly 2 months (58.15384615384615 days).

Some statistics for the networks are:

Densities

Time	Densities
period 0	0.1903112155020552
period 1	0.12237228420434527
period 2	0.08620082207868468
period 3	0.11374045801526718
period 4	0.030651790957134467
period 5	0.033294186729301234
period 6	0.020963006459189665

Change

Time	Change
period 0 / 1	0.5146446713030797
period 1 / 2	0.5466301605038167
period2 / 3	0.33860487310217624
period 3 / 4	0.6502059053707042
period 4 / 5	0.28102191580190716
period 5 / 6	0.32237759319753106

3 Models

3.1 Goal

The goal of the models is to see whether given a set of feature at t_0 you can predict whether there will be an edge (in this case a friendship) between nodes at point t_1.

3.2 General approach

In order to run a machine learning algorithm to recommend edges (which would take two nodes, a source and a candidate destination, and generate a score measuring the likelihood that the source would follow the destination), it's necessary to prune the set of candidates to run the algorithm on. The approach was to calculate a personalized PageRank around each source node.

I then only considered the 25 highest scoring candidate nodes for the next step. After pruning the set of candidate destination nodes to a more feasible level, I fed pairs of (source, destination) nodes into a machine learning model. In this case a random forest model.

3.3 Base

Features:

- Time spent together with another node

To add:

- Adamic Adar scores
- Network similarity measures

3.4 Social

Base model + additional features:

- Spatial-triadic closure
- Potential spatial-triadic closure
- Number of other people present

3.5 Time

Base model + additional features:

- Mode of the hour of the day
- Mode of the day of the week
- Mode of the hour of the week

3.6 Place

Base model + additional features:

- Met at home
- Met at university
- Met at third place

3.7 Social & Time

Features: Combination of social and time model

3.8 Place & Time

Features: Combination of place and time model

3.9 Social & Place

Features: Combination of social and place model

3.10 Social & Place & Time

Features: Combination of all features

4 Outcomes

At the moment I split the data into 10 bins and use 9 bins to train a random forest model and test it on the 10th bin.

As a measure the Mean Average Precision of the predicted links was used. The outcomes for the first time-step are as follows:

Model	MAP
base	0.8429072629997929
time	0.9186540879105468
social	0.920801185866502
place	0.842442824165898
social & time	0.9403508668997127
place & time	0.9206597144806115
social & place	0.9203355958501043
social & place & time	0.9405911908613162

5 Next steps

- Do the analysis for all time steps
- Add additional features
- Assess the importance of the features
- Update MAP measure

6 Open questions

- How to best define friendship?
- How many bins for testing?
- What time periods to consider? All or just the first few as they are densest?