Space, Time, and Sociability: Predicting Future Interactions

Christoph Stich

November 3, 2015

1 Introduction

Researchers have long studied the relationship between between space, time, and social structure as space, time and the social realm are intrinsically linked. For the most part people do not aimlessly amble, nor is happenstance usually the reason people spend time at places. People commute to work every day, they meet their friends at a bar after work, or they go on a date with their partner.

With the increased dissemination of GPS, mobile technologies, and social networks new avenues for research have opened up in recent years. For example, Backstrom et al ([?]) have found that the probability of friendship with a person decreases with distance. Scellato et al. ([?]) have studied the properties of location-based social networks and found that about 40% of all links in location-based social networks are shorter than 100km. Others ([?] and [?]) used the social and spatial properties of location-based social networks to propose a link-prediction model. While Brown et al. ([?]) developed a model for the evolution of city-wide location-based social networks, it remains unclear whether the qualities of a place itself fosters tie formation, or the fact that friends tend to meet at specific—more "social"—places.

Furthermore, Backstrom et al. ([?]) utilize the relationship between various geographic features and friendship the location of an individual from a sparse set of known user locations using the relationship between geography and friendship. Wang et al. ([?]) discover that the more similar two individuals are in their mobility the closer they are in the social network.

Last but not least, De Domenico et al ([?]) have used the mobility data of friends to consequentially improve user movement prediction, while Cho et al. ([?]) have built a mobility model incorporating both periodic movement of individuals as well as travel due to the social network structure. The exact interplay of the social structure and the human mobility patterns remains however unclear.

A number of important question however remain unstudied. For example, can we use chronological and spatial information to predict future interactions. So far work that has taken geographic features into account has focused on

relative static social structure and not interactions (quotes), whereas predicting interactions (Yang 2013) has not incorporated geographic information.

While the importance of time and space for social processes has been acknowledged by [the above cited people] () it is unclear whether place in itself has genuine predictive power or is simply confounded with the geographic embeddedness of social interactions. The unaddressed question is whether place actually matters or is it actually the people you happen to meet at different places.

Another open question is whether the type of relationship between nodes has any influence on the predictability of interactions. One could assume that meetings between colleagues are highly predictable as they both share a specific physical location—work—that both visit with high regularity. On the other hand one could assume that meeting your social ties is driven by a much more complex process and thus appear to be less regular.

To address those questions, we propose to phrase the question of whether two people will meet in a given time period as a link-prediction problem in a social interaction graph. We create a global link-prediction algorithm to assess the predictive power of time, place, and various social features have for different modeling scenarios. The rest of the paper is organized as follows: We formally define our problem in section 2. Section 3 describes the two datasets we use in our paper as well as explores our data in regards to the interaction between spatial and social variables. Section 4 details the setup for our prediction task, whereas section 5 discusses our findings. Last but not least we discuss the implications of our findings in section 6.

2 Problem Definition

2.1 Social interaction graph

We phrase the problem of predicting interaction as a link prediction problem in a time-varying, labeled network G that represents interactions. We define interaction any physical proximity measured by a strong bluetooth measurement. We use -80 [some unit] as Vedran () has shown this to be a reliable cut-off value for close and unobstructed physical proximity [expand a bit]. For each timepoint t we build a graph $G_t = (V_t, E_t)$, where V_t are the set of students at time t and E_t the set of links between them. Each edge $e \in E$ has also an associated weight $w(e) \in \mathbb{Z}$ representing the amount of interaction between any two nodes $u, v \in V$ in the previous period $\triangle T$. Let W_t be the distribution of assigned weights at a given timepoint t and W be the distribution of all weights over all possible timepoints. We assign a label $L(w(e)) \in \{0, 1, 2, 3\}$ to each $e \in E_t$. The labels $L(e)$ are based on the amount of interaction in the preceding time period and the cut-off values for each label are defined in relation to the quartils of

W. Let $L(e)$ now be $\begin{cases} 0: & m = 0 \\ 1: & 0 < m < Q_1(W) \\ 2: & Q_1(W) < m < Q_3(W) \\ 3: & Q_3(W) < m \end{cases}$ and m the observed time of interaction between any two $u, v \in V$ during $\triangle T$. These particular labels were chosen to be able to distinguish between *weak ties* ($l = 1$), *ties* ($l = 2$), and *strong ties* ($l = 3$). [I have to write something here about this]

2.2 Link-prediction

In a human interaction network $G_t = (V_t, E_t)$, the multi-class link prediction task is to predict $L(e)$ at time $t + \triangle t$, where $e(u, v) \notin E_t$ and $u, v \in V_t$.

Intuitively, we are trying to predict who will meet whom and for who long during period $\triangle t$. Formulating the problem this way has the advantage of including link dissolution—a not well studied problem in link-prediction (quote survey)—quite naturally in the problem definition. This is equivalent to predicting the labeled, network structure of $G_{t+\triangle t}$.

3 [Exploring the data]

3.1 Datasets

Our data consists of two datasets. [Write something about the datasets and how they were collected]

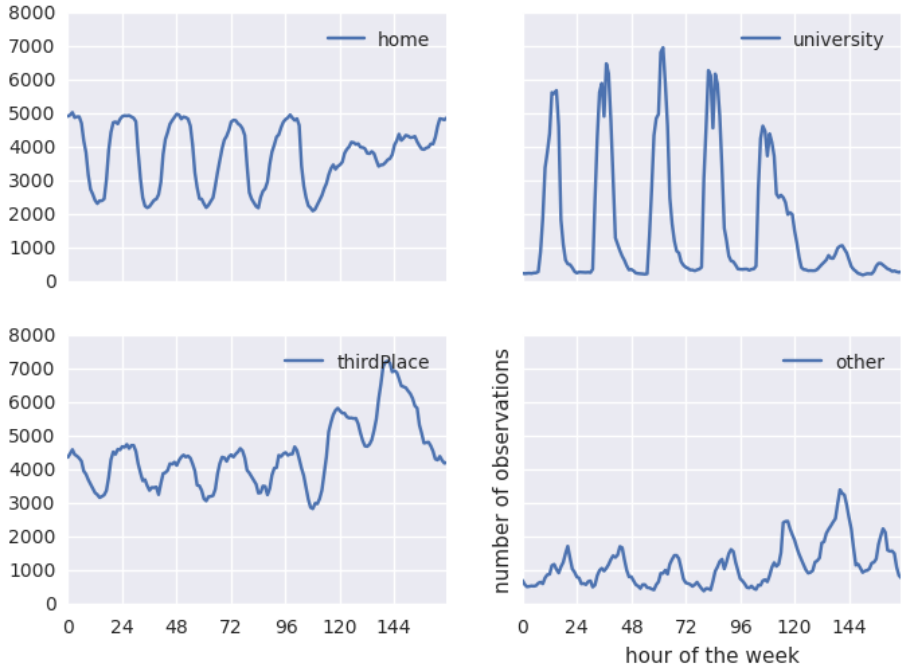
3.2 Descriptive parts

Because the literature reliably suggests a link between geography, time, and the social realm, we first investigate our data to check whether we can observe a relationship between those factors. If there is a relationship between space, time, and social interactions we should be able to observe differences in how people visit those places and in how they socialize at different geographic settings.

3.2.1 Geographic contexts

Based on Oldenburg's seminal paper () we develop a definition of different geographic contexts for our study, whose influence on link-prediction we are interested in. [write a short paragraph about oldenburg] Analogous to Oldenburg we distinguish between several different geographic settings a student can be in: the *home,* the *university,* a *third place*, and *other*. We infer the home location for each student by clustering all his or her location measurements between 11PM and 4AM using DBSCAN () into the set of clusters C. We then select $max(|c|), c \in C$ as a student's home location. For assigning students to the *university* context we mapped the campus of their university and checked whether students where within 50 meters of the campus. To infer the *third places* for

Figure 1: Weekly aggregated activity pattern



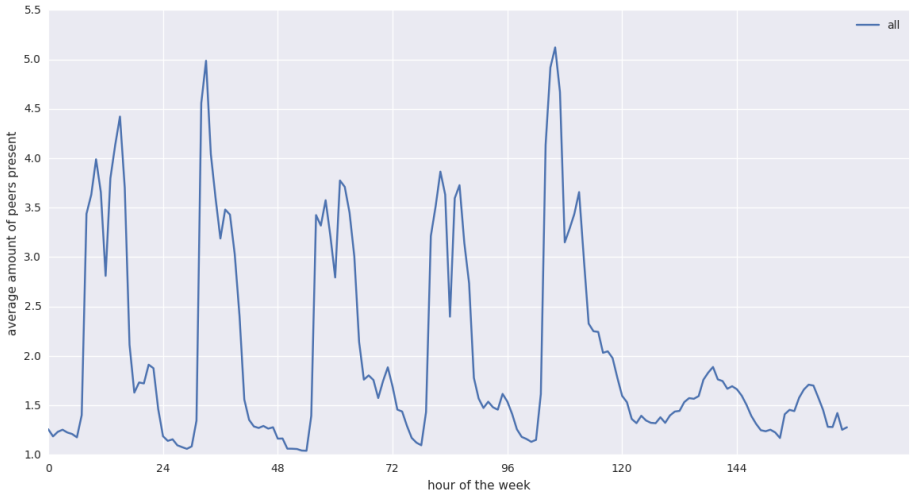
each student we first construct the set of all the stop locations a student would visit L_{stop}. For each $l \in L_{stop}$ we can also observe the amount of time *t(l)* a student spends there. We then define a *third place* as any $l \in L_{stop}$that fulfills the following inequality: $t(l)/\sum t(l_i) \geq 0.1$ and is not either *home* or *university*. A *third place* is thus any location where a student spends at least 10% of his time that is neither *home* nor *university*. Lastly any other $l \in L_{stop}$ is classified as *other*.

3.2.2 Activity pattern and social interactions

Looking at the aggregated weekly activity pattern of all students one can [say something about the figure]. First, we observe that during weekdays the two clearly dominating settings are *home* or *university* and the pattern is remarkably similar for all weekdays. Second, during weekends the dominant setting is *third Places* while almost no one visits the *university*. Also people explore a lot of *other* locations during the weekend as well. Thus, the behavior of students is consequentially different on weekdays than from weekends. Thus, it seems to safe to conclude that there is indeed a relationship between geographic setting and time as the observed behavior is too regular to be product of chance.

When looking at when people meet as in figure **??** one can clearly distinguish between weekdays and weekends and between days and nights. However, when

Figure 2: $\overline{x}_{peers}$



looking at the different geographic contexts the picture is less clear. Figure 3 shows the absolute deviation from the average amount of peers present for each geographic setting.

While one can spot differences their interpretation is less straight forward and one cannot determine whether the differences are due to chance. Lag plots are often used to determine whether a series is random or not (find a quote). An analysis of the lags in figure X reveals that the variations are not random as there are clearly linear patterns visible for all geographic contexts

If one looks at the the average edge life of each interaction between students for the different settings one can see that students interact for longer at home, at a third place, and at an other place then at university. At the university students interact for roughly 90 minutes, whereas the interactions at the other settings are significantly longer. Also the length of interactions is dependent on the time of the day. To offer a blunt interpretation: People don't interact when they are sleeping.

We thus conclude that are differences in how, when, for how long, and whom people meet and interact with in our data. Furthermore, the activities and interaction of the students are not random but follow identifiable patterns. Students socialize more at home and at places that are important to them, they do not visit the university during the weekend, and are more likely to be out during the weekend than at home. Taken at it's face value those findings are rather unexciting, but we will use those differences for predicting who will meet whom in the next section.

Figure 3: Deviation from $\overline{x}_{peers}$

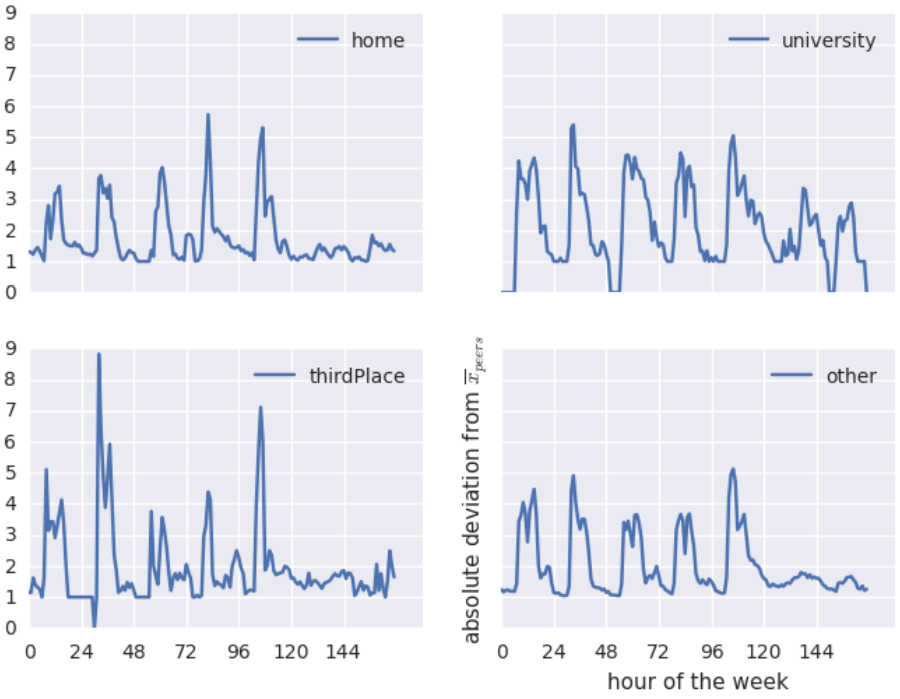


Figure 4: Lag plots deviations of $\overline{x}_{peers}$

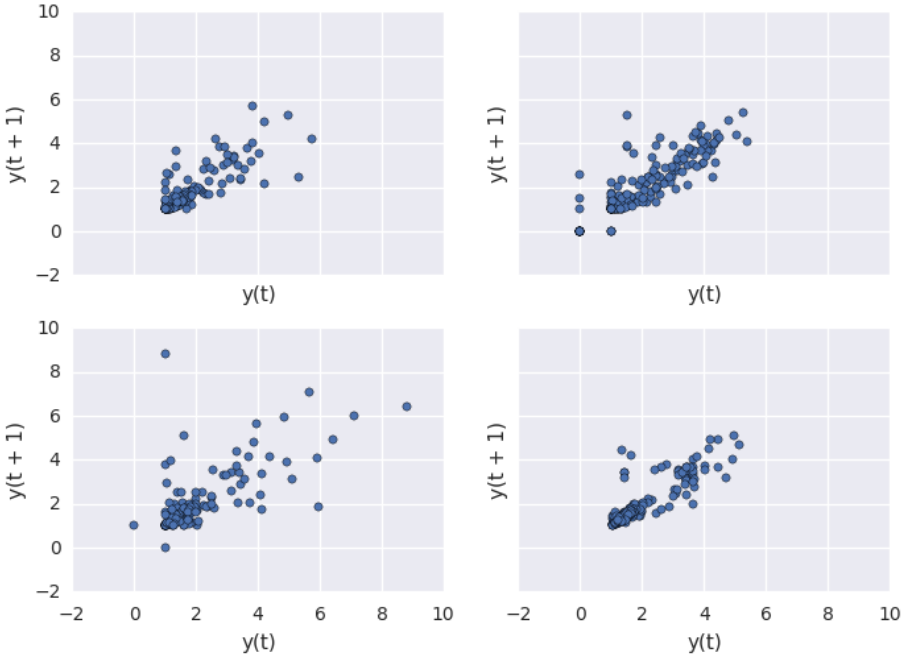
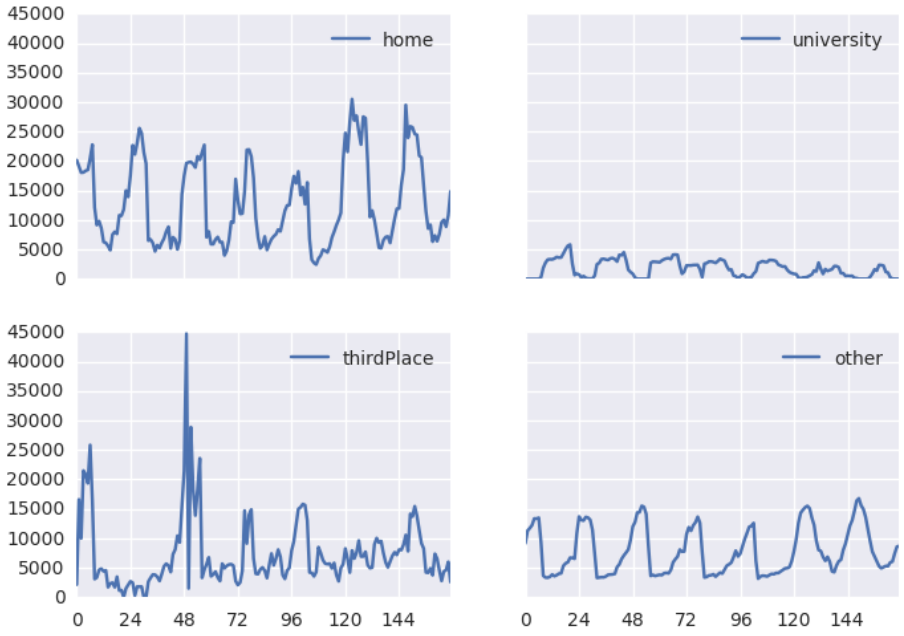


Figure 5: $\overline{x}_{edge\ Life}$



4 Link prediction

Random Forests have been consistently shown to perform well in link-prediction tasks (for example X and Y) and we thus opt to use them for our prediction task as well. We are particularly interested in the drivers of social interaction or in other words what set of features gives us the best prediction and to a lesser extent in evaluating different classifiers for the prediction task at hand.

4.1 Choosing $\triangle T$ and $\triangle t$

As Yang et al () have shown setting the length of $\triangle T$ and $\triangle t$ has an impact on the performance of the resulting link prediction. Yang et al () have proposed to use the time series of the density of the network as a guide for selecting $\triangle T$ and we mostly follow their approach here. When looking at the density time series (figure density time series) one can clearly identify a weekly pattern, but also seasonal effects. Periods of low density either coincide with holidays or with exam periods. We thus opt to use a $\triangle T$ of 14 days as this is the shortest time-period that captures the weekly periodicity as well as mitigates the negative effects of seasonality. Furthermore, we keep $\triangle T$ fixed in order to be able to compare the performance of the algorithm at different time-points througout our experiment.

Contrary to Yang et al () we did not opt to use the average duration of an interaction as $\triangle t$. While the average length of interaction is around 806 seconds, we chose to instead use several hours as$\triangle t$ ($\triangle t = 6h$). As this allows us to predict whether there will be an edge between u,v e V, but also to distinguish between the strength of the tie over the course of a longer time interval. This is important as we are much more interested in the drivers of interaction than in correctly predicting chance encounters. A shorter $\triangle t$ would not allow us to predict the length and thus the nature of the interaction. Furthermore, compare the entropy of different length of encounters in table X. We can see that longer encounters are have a consequentially lower entropy value than shorter ones, the ones we are most interested in. Furthermore, encounters of an undefined length make up around 0.49[check this figure on the whole dataset] of all encounters, i.e. encounters that appear in only sampling interval of the smartphone and we thus cannot assign a meaningful duration.

4.2 Search space

Usually researchers have restricted their search space for new ties as there are almost N potential candidates in sparse social networks. The complexity of any algorithm that searches in an unrestricted space is thus O(N^2). Common ways to deal with this are either considering only the set of friends of friends, "place-friends" (), "mobility" friends () as potential candidates for new ties. However, our network is small enough that is still computationally feasible to consider all possible pairs of nodes. Furthermore, we can observe a lot of change in the structure of the graph between time points (figure X) and restricting the search

space would exclude several potential candidates at each time step. We have to admit thought that our approach thus does not scale well to much larger datasets.

4.3 Feature vectors

4.3.1 Baseline features

As our baseline features for all subsequent models we include *recency*, the amount of elapsed time since the last meeting, *activeness,* how often two nodes interacted (Quotes from the Yang paper) and how much time spent two nodes spent together during the training period.

4.3.2 Context features

We also include several features pertaining to the setting wherein two nodes meet. These can be split into features relating to time, space, and the social realm. The time related features pertain to capture weekly behavioral patterns. Let M be the set of all meetings between two nodes u, v in the training period. We then include a *vector(hour-of-day(M))*, *vector((hour-of-week(M))*, *vector((day-of-week(M))*.

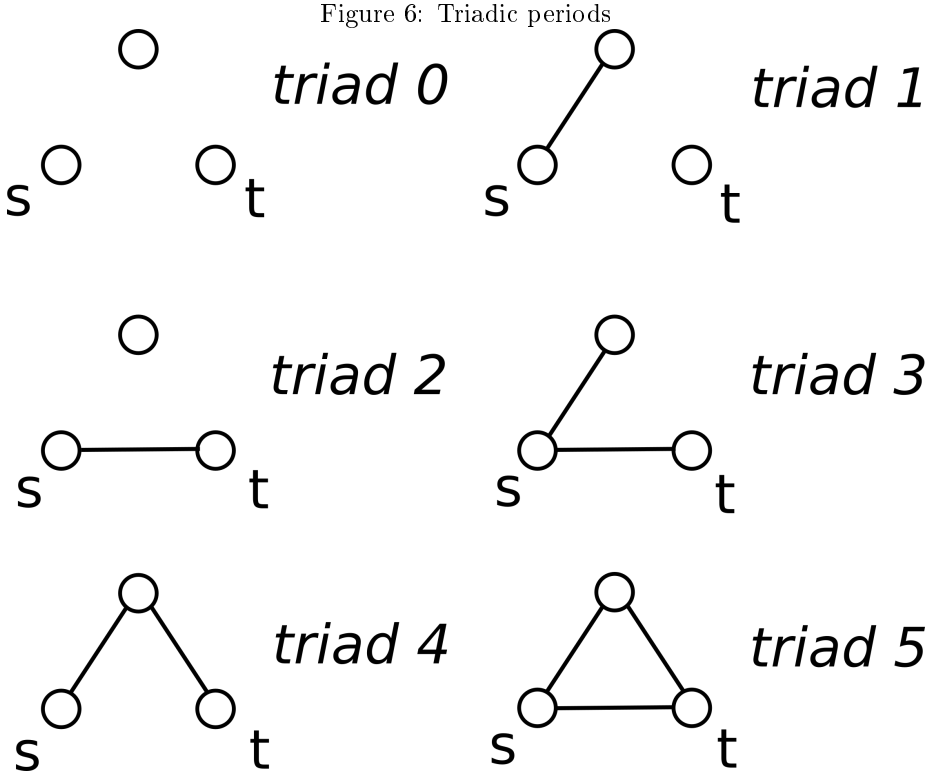
We also include *min(place entropy)* (quote) of the meetings as we reason that there is a difference in whether two people meet at a place a lot of people visit and thus with high place entropy or at "quieter" place with low place entropy. Or in other words, if two student meet at the university then this probably does not tell us that much as a lot of people are meeting there, but if two people meet at their respective homes then this is a much more unlikely and thus noteworthy event.

We also infer the *relative importance* of each venue for each user by measuring the amount of time a user spends there and consequently ranking them. We then include the *max(relative importance)* of a meeting between any $u, v \in N$.

Let $context(M)$ be the function that counts the amount of time two nodes u, v have spent together at the different geographic contexts (as defined in section something). We then include the vector(context(M)) as a feature as well. The reasoning being that by measuring the amount of time two nodes spent together in different geographic settings might allow us to gauge the quality of their relationship. For example, if two nodes only ever meet at the university they are likely just colleagues, but if they meet in other settings as well they might have a stronger relationship.

But it is not only the physical qualities of the place that might influence the pattern of interactions but also the social setting an interaction occurs. If two students meet at the university during a course this does is not extraordinary, but if two students meet alone on the campus there is a higher likelihood that they are socializing. Let now P be the distribution of the number of other people from the study that present when two nodes u and v meet. We then include *min(P)* as well as *avg(P)* as features.

Figure 6: Triadic periods



Triadic-closure that is in social network the process that friends of friends are likely to become my friends as well has been known to play a significant role in network formation (find a quote). Yang et al (quote) have used the "triadic periods" successfully as features for predicting interactions before. We build upon their work and adapt their metric for our problem. The main idea is to count the different possible arrangements of triads in the interaction graph, or in other words the different possible configurations of collocations. Figure [X] shows the possible arrangements of collocation triads (excluding symmetric triads).

4.3.3 Network features

We also include several features that are based on the network topology of the interaction graph and have been used in link-prediction problems before. In particular we include Adamic Adar (quote), the Jaccard coefficient (quote), preferential attachment (quote), ressource allocation (quote), and random walk with restarts (quote).

4.4 Null Model

As a benchmark to test our predictions against we also developed a null model for a time-evolving weighted interaction graph with dissolving ties. The null model asserts that change between time-points in G is happening randomly, while it adheres to the true amount of change of the graph between time-points. "True change" in our case means created ties ($E_{t+\triangle t} \backslash E_t$) as well as dissolved ties ($E_{t+\triangle t} \backslash E_t$) for each class. Where we take the probability that a tie changes classes between time-points - $P(x_{t+\triangle t}|y_t)$ where $x, y \in \{0, 1, 2, 3\}$ - from the observation of the actual change between G_t and $G_{t+\triangle t}$.

4.5 Experimental Setup

We developed and tested it our model on our smaller dataset before running the "completed" model on our second, larger and independent dataset. This way we avoid biasing ourselves and developing our model to fit our data. Only after we fully developed our model, did we proceed to re-run our algorithm on the second, bigger dataset. [I need a quote for this and add some words].

5 Findings

5.1 [Findings of the general model]

5.2 [Multi-class vs single-class case]

5.3 [Findings for the different models]

5.4 [Performance of the node-only model]

Sometimes one however might not have access to the whole network and one might only be in possession of more or less isolated node level data. Consequently one is unable to calculate or reliably estimate the network features we describe in section [something]. We simulate such a scenario by building another model that only incorporates data on a node-level. [We can see something and we do something]

[Probably we can see that the model works okay and that for node-level information the setting information is more important]

5.5 [Performance of several different T lengths]

5.6 [network prediction performance]

6 Conclusion/Discussion

[expand on node-level prediction: Potentially having separate models for each user]