Randomization Tests for Distinguishing Social Influence and Homophily Effects

Timothy La Fond and Jennifer Neville
Computer Science Department
Purdue University
West Lafayette, IN 47907
[tlafond|neville]@cs.purdue.edu

ABSTRACT

Relational autocorrelation is ubiquitous in relational domains. This observed correlation between class labels of linked instances in a network (e.g., two friends are more likely to share political beliefs than two randomly selected people) can be due to the effects of two different social processes. If *social influence* effects are present, instances are likely to change their attributes to conform to their neighbor values. If *homophily* effects are present, instances are likely to link to other individuals with similar attribute values. Both these effects will result in autocorrelated attribute values. When analyzing static relational networks it is impossible to determine how much of the observed correlation is due each of these factors. However, the recent surge of interest in social networks has increased the availability of dynamic network data. In this paper, we present a randomization technique for temporal network data where the attributes and links change over time. Given data from two time steps, we measure the gain in correlation and assess whether a significant portion of this gain is due to influence and/or homophily. We demonstrate the efficacy of our method on semi-synthetic data and then apply the method to a real-world social networks dataset, showing the impact of both influence and homophily effects.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Design

Keywords

Social networks, randomization, homophily, social influence

1. INTRODUCTION

Autocorrelation is a common characteristic of relational and social network datasets, which refers to a statistical dependency between the values of the same variable on related entities. For example, friends are more likely to share political views than randomly selected pairs of individuals. The presence of autocorrelation offers a unique opportunity to improve predictive models because inferences about one object can be used to improve inferences about related objects. Indeed, recent work in relational learning has exploited this property in the development of *collective inference* models, which can make more accurate predictions by jointly inferring class label values throughout a network (see e.g., [5, 18, 24]). In addition, the gains that collective model achieve over conditional models (which reason about each instance independently) increase as autocorrelation levels increase in the data [10].

A number of widely occurring phenomena give rise to autocorrelation dependencies. Social phenomena, including social influence [13], diffusion processes [7], and the principle of homophily [15], can cause autocorrelated observations through their influence on social interactions that govern the data generation process. Alternatively, a hidden condition or event, whose influence is correlated among instances that are closely located in time or space, can produce autocorrelated observations through joint influence on link and attribute changes [17, 2].

A key question for understanding and exploiting behavior in social network domains is to determine the root *cause* of observed autocorrelation. Since autocorrelation is the primary motivation to use relational and network models over conventional machine learning techniques, it stands to reason that a better understanding of the causes of autocorrelation will inform the development of improved models and learning algorithms. For example, although previous work in relational learning and statistical network analysis has focused primarily on static graphs, recent efforts have turned to the analysis of dynamic networks and development of temporally-evolving models (e.g., [9, 21]). In order to deal with the enormous increase in dimensionality associated with modeling *both* temporal and relational dependencies, these methods restrict the set of dependencies that they consider (e.g., through choice of model form). The ability to accurately distinguish which temporal-relational patterns (e.g., homophily) occur in real-world datasets will ensure that researchers can include the most promising set of dependencies in their restricted set of patterns.

Research in social psychology and sociology has developed two main theories of social processes that can indicate why autocorrelation is often observed in social systems. *Social influence* refers to processes in which interactions with others causes individuals to conform (e.g., people change their attitudes to be more similar to their friends). *Homophily* refers to processes of *social selection*, where individuals are more likely to form ties with “similar” individuals (e.g., people choose to be friends with people who share their beliefs). Both homophily and social influence can produce autocorre-

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

lation, since their outcome results in linked individual sharing attribute values.

In this work we focus on the task of differentiating between influence and homophily effects and determining, from the observed autocorrelation dependencies, whether the effects are *significant*. Recently, there have been a number of empirical studies that investigate (and model) either social influence or homophily effects in real-world datasets (e.g., [4, 22, 6]. However, these efforts have focused primarily on demonstrating the presence of homophily and influence—they do not provide the means to estimate effects sizes from data or determine whether the effects are statistically significant. Exceptions include the work of Snijders et al. [23], Anagnostopoulos et al. [1], and Aral et al. [3]. Snijders et al. [23] develop a time-evolving exponential random graph model that can represent homophily and influence effects. Their method support hypothesis tests for each effect, but the applicability of the approach is limited by the suitability of the model form (e.g., random graph model, Markov assumption). On the other hand, the recent work of Anagnostopoulos et al. [1] presents a model-free approach to assessing influence effects with randomization tests. The limitation of their framework, however, is an assumption that that the network structure (i.e., links) does not change over time, thus they cannot distinguish homophily effects. Aral et al. [3] correct this issue with a development of matched sample estimation framework that accounts for homophily effects, but the method uses additional node behaviors and characteristics in the matching process, so it will have limited applicability in data with few observed attributes and/or time steps.

In this paper, we outline a more general randomization framework for datasets where *both* attribute values and links change over time, where changes can consist of either additions or deletions. Our aim is to determine the significance of each effect and to *distinguish* the contribution of influence and homophily effects. We outline a randomization test based on randomization of *action choices*. We consider the *gain* in correlation over one time step in the graph and assess the amount of gain that is due to each of the effects. The randomization procedure then produce an empirical sampling distribution of expected gains under the null hypothesis (that there is no influence and/or homophily effect) and if the observed gain is greater than expected under the null, we can conclude there is a significant influence/homophily effect.

We evaluate our proposed method on semi-synthetic social network data, showing that the test has low Type I error (i.e., it does not incorrectly conclude there is an effect when in fact the data are random) and high power when the data exhibit sufficient change over time (i.e., they correctly conclude there is an effect when there is one). We then apply our method to a real-world dataset to investigate the aspects of observed autocorrelation. Our analysis of a public university Facebook network shows that autocorrelation in group memberships is due to significant influence and homophily effects, yet different groups exhibit different types of behavior.

2. PROBLEM DEFINITION

In this work, we consider relational data represented as an undirected, attributed graph $G = (V, E)$, with V (nodes) representing objects and E (edges) representing relationships. The nodes V represent objects in the data (e.g., people) and the edges E represent relationships (e.g., friendships) between pairs of objects ($e_{ij} : v_i$ and v_j are friends). Each node $v \in V$ and has a number of associated attributes $\mathbf{X^v} = (X_1^v, ..., X_m^v)$ (e.g., age, gender).

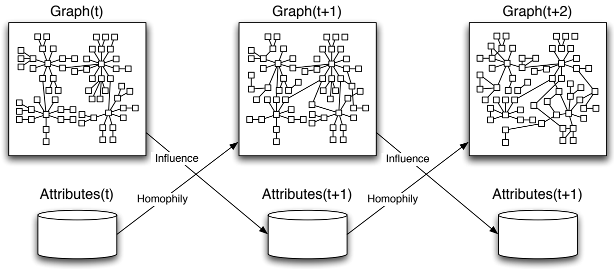


Figure 1: Illustration of homophily and influence affect on attributes and links over time.

We assume that both the attributes and links may vary over time. First, attribute values may change at each time step t: $\mathbf{X_t} = \{\mathbf{X_t^v}\} = \{(X_{1t}^v, ..., X_{mt}^v)\}$. Second, relationships may change at each time step. This results in a different data graph $G_t = (V, E_t)$ for each time step t, where the nodes remain constant but the edge set may vary (i.e., $E_t \neq E_{t'}$ for some t, t').

Figure 1 illustrates *influence* and *homophily* dependencies. If there is a significant influence effect then we expect the attribute values in $t + 1$ will depend on the link structure in t. On the other hand, if there is a significant homophily effect then we expect the link structure in $t + 1$ will depend on the attributes in t.

If either influence or homophily effects are present in the data, the data will exhibit *relational autocorrelation* at any given time step t. Relational autocorrelation refers to a statistical dependency between values of the same variable on related objects—it involves a set of *related* instance pairs, a variable X defined on the nodes in the pairs, and it corresponds to the correlation between the values of X on pairs of related instances. Any traditional measure of association, such as Pearson's correlation coefficient or information gain, can be used to assess the association between these pairs of values of X. In this work, we use the chi-square statistic.

Definition 1. **Relational Autocorrelation**
Let $P_R = \{(v_i, v_j) : e_{ij} \in E\}$ be a set of *related* instance pairs in G. Let X be a binary attribute defined on the nodes V. Then we compute the relational autocorrelation of X in G with the following contingency table T:

	$X^i = X^j = x$	$\neg(X^i = X^j = x)$
$(v_i, v_j) \in P_R$	a	b
$(v_i, v_j) \notin P_R$	c	d

We define *relational autocorrelation* as the chi-square statistic that is computed from T (with *dof*=1):

$$C(X, G) = \chi^2 = \frac{(ad - cb)^2 \cdot N}{(a + b)(c + d)(b + d)(a + c)}$$

where $N = a + b + c + d$, is the total count of all cells in T.

The first column of the contingency table counts pairs of nodes that both have the same value for attribute X. The second column counts pairs of nodes that do not match on

X. The first row counts pairs of nodes that are related in G. The second row counts pairs of nodes that are not linked in G. Note that this contingency table encompasses every possible combination of nodes in the graph and thus has a stable size even when the total number of links change in the graph (i.e., it doesn't depend on the size of E).

To measure the autocorrelation between attributes and relationships at time t, we compute the chi-square statistic $C(X_t, G_t)$ from the graph G_t using the attribute values in X_t. Using a similar table for the attributes and links in $t+1$, we can compute $C(X_{t+1}, G_{t+1})$.

Now that we have a method of computing the correlation at any time step, we can can use correlation *gain* to assess the effects of homophily and influence, where the observed *correlation gain* G from t to $t+1$ is:

$$gain(t, t+1) = C(X_{t+1}, G_{t+1}) - C(X_t, G_t)$$

The gain in correlation from one time step to the next can be due to: (1) homophily gains due changes in the graph structure in $t+1$, or (2) influence gains due to changes in attributes in $t+1$. To show this, we will define influence and homophily and show how they impact the chi-square statistic from one time step to the next.

"Homophily" is typically used to refer to the general tendency of people to associate with similar others (see e.g., [15])—we operationalize this in the following way to investigate whether attribute similarity influences choice of friends (i.e., are friendships formed based on a pair's attribute similarity).

Definition 2. **Homophily**
Let X_t and X_{t+1} be the attribute values at time t and $t+1$ respectively. Let $P_{R(t)}$ and $P_{R(t+1)}$ be the related pairs at time t and $t+1$ respectively. Let L^{ij}_{t+1} refer to the case when pair (v_i, v_j) form a link at time $t+1$ (i.e., $(v_i, v_j) \notin P_{R(t)}$ and $(v_i, v_j) \in P_{R(t+1)}$). Let U^{ij}_{t+1} refer to the case when pair (v_i, v_j) drops a link at time $t+1$ (i.e., $(v_i, v_j) \in P_{R(t)}$ and $(v_i, v_j) \notin P_{R(t+1)}$). Then a dataset exhibits *homophily* if the following hold:

$$p(L^{ij}_{t+1}|(X^i_t = X^j_t = x)) > p(L^{ij}_{t+1}|\neg(X^i_t = X^j_t = x))$$
$$p(U^{ij}_{t+1}|\neg(X^i_t = X^j_t = x)) > p(U^{ij}_{t+1}|(X^i_t = X^j_t = x))$$

In other words, the probability of link formation over time (from t to $t+1$) is higher for pairs with matching attribute values and the probability of link dissolution over time is higher for non-matching pairs.

"Social influence" is typically used to refer to the general case of a person's behavior being influenced by others (see e.g., [14])—we operationalize this in the following way to investigate whether a person's friends influence their intrinsic attributes (i.e., are attribute values changed to match one's friends).

Definition 3. **Social Influence**
Let X_t and X_{t+1} be the attribute values at time t and $t+1$ respectively. Let $P_{R(t)}$ and $P_{R(t+1)}$ be the related pairs at time t and $t+1$ respectively. Let A^{ij}_{t+1} refer to the case when pair (v_i, v_j) change their attribute values to agree at time $t+1$ (i.e., $\neg(X^i_t = X^j_t = x)$ and $(X^i_{t+1} = X^j_{t+1} = x)$). Let D^{ij}_{t+1} refer to the case when pair (v_i, v_j) change their attribute values to diverge at time $t+1$ (i.e., $(X^i_t = X^j_t = x)$ and $\neg(X^i_{t+1} = X^j_{t+1} = x)$). Then a dataset exhibits *social influence* if the following hold:

$$p(A^{ij}_{t+1}|(v_i, v_j) \in P_R(t)) > p(A^{ij}_{t+1}|(v_i, v_j) \notin P_R(t))$$
$$p(D^{ij}_{t+1}|(v_i, v_j) \notin P_R(t)) > p(D^{ij}_{t+1}|(v_i, v_j) \in P_R(t))$$

In other words, the probability of agreement over time (from t to $t+1$) is higher for related pairs and the probability of disagreement over time is higher for unrelated pairs.

Given these definitions we can show that homophily and influence will result in a correlation gain over time.

THEOREM 1. ***Influence Gain***
Let X_t and X_{t+1} be attribute values at time t and $t+1$ respectively and let $P_{R(t)}$ be the related pairs at time t. Let $k = |A^{ij}_{t+1}|$ be the number of agreements from t to $t+1$. Let $m = |D^{ij}_{t+1}|$ be the number of disagreements from t to $t+1$ and let $k = m$. Then if an influence effect is present in the data, the autocorrelation will increase when we consider the attribute changes from time t to time $t+1$:

$$C(X_{t+1}, G_t) > C(X_t, G_t)$$

The proof of this theorem is included in Appendix A.

THEOREM 2. ***Homophily Gain***
Let $P_{R(t)}$ and $P_{R(t+1)}$ be the set of related nodes at time t and $t+1$ respectively and let X_t be the attributes at time t. Let $k = |L^{ij}_{t+1}|$ be the number of link additions from t to $t+1$. Let $m = |U^{ij}_{t+1}|$ be the number of link dissolutions from t to $t+1$ and let $k = m$. Then if a homophily effect is present in the data, the autocorrelation will increase when we consider the link changes from time t to time $t+1$:

$$C(X_t, G_{t+1}) > C(X_t, G_t)$$

The proof of this theorem follows the same form and argument as for Theorem 1.

Now that we have illustrated the connection between gains in autocorrelation and influence/homophily, we can define the correlation decomposition problem as follows. Given an observed network change over two time steps, G_t, $\mathbf{X_t}$, G_{t+1}, $\mathbf{X_{t+1}}$, determine whether (1) the attribute changes from t to $t+1$ exhibit a significant amount of *influence*, and (2) the link changes from t to $t+1$ exhibit a significant amount of *homophily*. In section 4, we will outline a novel randomization test to separate these effects and assess their significance.

3. RELATED WORK

Current approaches relevant to this work fall into three categories: empirical investigation and modeling of social influence and homophily effects, significance tests for relational and social network data, and modeling techniques for distinguishing homophily and influence effects.

Researchers in social psychology and sociology have studied social influence and homophily for much of the past forty years (see [14] for an extensive review). This work has focused primarily on developing theory about underlying psychological processes such as persuasion, conformity, assimilation, and selection. Experimental investigation is typically performed in smaller-scale laboratory environments, with individual or dyad-level analysis. Consequently, the modeling techniques do not need to model the interdependence among individuals, nor do they need to be applicable for large-scale networks of thousands of nodes.

Recently, with the surge of interest in online social networks and relational data, there has been a growing body of research in the data mining community that has focused on modeling network and behavior change over time. For example, Backstrom et al. [4] investigated the evolution of network structure and group membership in MySpace and LiveJournal and showed that homophily can be used to improve predictive models of group membership. Singla and Richardson [22] investigate the correlation between individual search topics among people that interact using instant messaging, and show that not only does a correlation exist but that it increases with the amount of time the users communicate. Crandell et al. [6] study the temporal evolution of link structure and attribute similarity in Wikipedia and propose a mathematical model that includes both influence and homophily effects to predict future behavior in the network.

The majority of this recent work is empirically-based—focused primarily on demonstrating the presence of homophily and influence in real-world data. The proposed methods and analysis techniques do not provide the means to estimate effects sizes from data or determine whether the effects are statistically significant.

There has been some work on developing significance tests for social network and relational domains, but this work has focused primarily on static networks, so the null models clearly limit the types of conclusions that can be drawn from the analysis. For example, Milo et al. [16] generate randomized network structures while holding node degree constant to assess whether subgraph motifs are observed a significantly higher frequency than would be expected due to random chance. Karrer et al. [12] consider the significance of community structure in the network, by using network perturbations to assess the variance of community patterns. Jensen et al. [11] propose attribute-based randomization tests to accurately assess the significance of feature association in the presence of relational autocorrelation. Eldardiry and Neville [8] develop resampling procedures for attributed networks that can also be used to assess variance of feature scores in static networks. In addition, work in sociology on exponential random graph models (see e.g., [20]) includes methods for determining the significance of parameter estimates learned from data (again on static networks).

The work most relevant to our proposed method is that of Snijders et al. [23], Anagnostopoulos et al. [1], and Aral et al. [3]. Snijders et al. [23] extend the exponential random graph models to incorporate time evolution (in both attributes and links) with a Markov model assumption (i.e., attribute/links at one time step depend only the previous time step). The method facilitates the inclusion of social influence and homophily dependencies in the model and generalized Neyman-Rao score tests, based on methods of moments estimators, are used to test hypotheses of the form: $H_0 : \theta_i = 0, H_1 : \theta_i \neq 0$. The main limitation of this approach is that it is model-based so the accuracy of hypothesis tests will be impacted by the suitability of the model form (e.g., random graph model, Markov assumption). In contrast to the work of Snijders et al. [23], our method is a data-driven, model-free approach based on randomization tests.

The recent work of Anagnostopoulos et al. [1] also outlines a randomization procedure for assessing whether an observational datasets exhibits a significant influence effect. However, their framework assumes that the network structure (i.e., links) does not change over time, thus they do not present a method for assessing whether a significant homophily effect is present as well. In particular, their timestamp shuffling test requires a longitudinal view of the evolution of the data. To estimate the effects of social influence, they compare the number of people the adopt a trait, given they have a neighbors with that trait already, to the number of people that do not adopt, given the same a neighbors with the trait. This calculation requires future knowledge about who will not adopt a trait. Otherwise, if the comparison is made using a single time step, the vast majority of users will not have adopted and the effect will lost in the noise. Furthermore, their method only considers data where attributes are added over time.

Concurrent work by Aral et al. [3] has corrected the main limitation of [1] in their development of a matched sample estimation framework, which accounts for homophily effects as well as influence. However, their method uses additional node behaviors and characteristics in the matching process, so it will have limited applicability in data with few observed attributes and/or time steps. In this work, we outline a general randomization framework where *both* attribute values and links change over time, changes can consist of additions or deletions, and focus on assessing *both* influence and homophily effects in data with few available time slices.

4. METHOD

Randomization tests are a model-free, computationally-intensive statistical technique for hypothesis testing [19]. The tests generate many replicates of an actual data set—typically called pseudosamples—and uses the pseudosamples to estimate a score distribution. Pseudosamples are generated by randomly reordering (or permuting) the values of one or more variables in an actual data set. A score is then calculated for each pseudosample, and the distribution of these randomized scores is used to estimate a sampling distribution for the score statistic under the null hypothesis. The value of the observed score on the original data is then compared to the distribution of scores on the randomized pseudosamples, and if it is significantly higher (or lower) than this distribution, the observed score will be deemed significant.

In contrast to conventional hypothesis tests, randomization tests make a relatively small number of assumptions about the data. For example, randomization tests make no assumptions about the form of the distributions from which variable values are drawn. In addition, they can be used to form sampling distributions for estimators whose precise statistical properties are not known. The key issue in developing a randomization test is to formulate an appropriate null hypothesis and permute the data in a way that accurately reflects the null hypothesis.

Tables 1-2 outline the specific significance tests that we use throughout this work. The significance tests determine whether an observed gain in autocorrelation is *significant* by using a randomization test to estimate an empirical sampling distribution of gains that would be expected if change in links (attributes) is random and thus not due to homophily (influence).

The empirical sampling distribution is estimated from the gains observed in *pseudosamples* generated by the random-

```
$HomophilySigTest(G_t, G_{t+1}, X_t, X_{t+1}, numIters, \alpha)$
$gains_R = \emptyset$
// compute original gain
$gain_O = C(X_t, G_{t+1}) - C(X_t, G_t)$
// randomize
For $iter$ in 1.. $numIters$
    $G'_{t+1} = Randomize(G_t, G_{t+1})$
    Compute $gain_r = C(X_t, G'_{t+1}) - C(X_t, G_t)$
    $gains_R = gains_R \cup \{gain_r\}$
// test significance of gain
If $gain_O > 1 - \frac{\alpha}{2}$ critical value of $gains_R$
    Return *significant/positive*
Else if $gain_O < \frac{\alpha}{2}$ critical value of $gains_R$
    Return *significant/negative*
Else
    Return *not significant*
```

Table 1: Homophily significance test method

```
$InfluenceSigTest(G_t, G_{t+1}, X_t, X_{t+1}, numIters, \alpha)$
$gains_R = \emptyset$
// compute original gain
$gain_O = C(X_{t+1}, G_t) - C(X_t, G_t)$
// randomize
For $iter$ in 1.. $numIters$
    $X'_{t+1} = Randomize(X_t, X_{t+1})$
    Compute $gain_r = C(X'_{t+1}, G_t) - C(X_t, G_t)$
    $gains_R = gains_R \cup \{gain_r\}$
// test significance of gain
If $gain_O > 1 - \frac{\alpha}{2}$ critical value of $gains_R$
    Return *significant/positive*
Else if $gain_O < \frac{\alpha}{2}$ critical value of $gains_R$
    Return *significant/negative*
Else
    Return *not significant*
```

Table 2: Influence significance test method

ization procedure. The method compares the observed gain value to the empirical sampling distribution, if the value is higher than $(1 - \frac{\alpha}{2})$% (or lower than $\frac{\alpha}{2}$%) of the scores observed in the randomized data, the gain is deemed to be *significant* and the null is rejected.

The gain in autocorrelation from one time step to the next can be due to: (1) homophily gains due to friend changes in $t + 1$, or (2) influence gains due to changes in attributes in $t + 1$. To separate the effects of influence and homophily, we define two different randomization tests to use as the *Randomize( )* method inside the significance test.

The key issue in developing a randomization test is to formulate an appropriate null hypothesis and permute the data in a way that accurately reflects the null hypothesis. We formulate three null hypotheses with respect to homophily and influence:

- H_0^H: link changes are random and are not due to attribute values in t (i.e., no homophily effect)
- H_0^I: attribute changes are random and are not due to friends in t (i.e., no social influence effect)
- H_0^F: both attribute and link changes are random (i.e., no homophily nor influence effect)

To identify possible permutations for these null hypotheses, we consider four types of data changes that can occur in the data from time t to time $t + 1$:

Edge additions: $\Delta_E^+ = \{e_{ij} \in E_{t+1} \wedge e_{ij} \notin E_t\}$

Edge deletions: $\Delta_E^- = \{e_{ij} \in E_t \wedge e_{ij} \notin E_{t+1}\}$

Attribute additions: $\Delta_X^+ = \{x^v \in X_{t+1} \wedge x^v \notin X_t\}$

Attribute deletions: $\Delta_X^- = \{x^v \in X_t \wedge x^v \notin X_{t+1}\}$

Note that attribute value changes can be easily modeled as an addition/deletion pair. Clearly, homophily will impact edge changes and influence will affect attribute changes.

For the null hypothesis concerning homophily (H_0^H), we want to randomize the edge changes to remove any association with attribute values in t. To do this, we can randomize the choice of edge target so that it does not depend on the attributes of the source node. For example, if node i adds a link to node j at time $t + 1$, then we can maintain the edge addition in $t + 1$ but randomize the choice of target node j to replace e_{ij} with $e_{ij'}$ so that any association of attribute similarity between i and j is destroyed. However, to ensure that the degree of j' remains the same after randomization, j' must have been part of an edge addition $e_{kj'}$ in the original set. The randomization procedure for edges can be thought of as swapping the endpoints of edge additions/deletions such that each node will have the same number of additions and deletions in the randomized set, but the partner of those links will have changed.

Randomization for the null hypothesis concerning influence (H_0^I) will follow a similar procedure by swapping attribute adoptions and abandonments between nodes, removing any influence of edges in t. If node i adds a attribute value x at time $t + 1$, then we can maintain the attribute addition in $t+1$ but randomize the choice of value to replace x with x' so that any similarity of the attribute value x with the attribute values of i's linked friends in t is destroyed.

We call the procedures based on this form of randomization *choice-based* methods, since they randomize the results of choices (attribute/link changes). Tables 3-4 outline the specifics of the choice-based method for H_0^H (homophily) and H_0^I (influence) respectively. We can combine the two methods, randomizing both the attribute and link changes to estimate a distribution for H_0^F.

However, calculating choice-based randomizations are nontrivial. A particular target edge or attribute can be selected for swapping only if it has not been selected before, and nodes cannot add edges or attributes if they had them in time step t, nor can they drop edges or attributes if they lack them in t. Given these sets of constraints, it may be difficult to find a valid random assignments (apart from the original), which is problematic for a test that depends on generating a distribution of random pseudosamples.

We address this issue by taking a greedy assignment approach. First, we collate the edge and attribute changes such that all additions and deletions for a node or attribute can be decided at once. Then, we sort the nodes and attributes from those with the least number of random options to those with the largest number of random options. Random options here refers to the amount of freedom the node has when selecting additions or deletions and is given by the number of available selections minus the number of assignments needed. This value can be calculated for edge

```
$Randomize^{hom}_{choice}(G_t, G_{t+1})$
$\Delta^+_E = \{e_{ij} \in E_{t+1} \wedge e_{ij} \notin E_t\}$ (added links in $t+1$)
$\Delta^-_E = \{e_{ij} \in E_t \wedge e_{ij} \notin E_{t+1}\}$ (dropped links in $t+1$)
// targets for random selections
$T^+ = \Delta^+_E$
$T^- = \Delta^-_E$
// randomize
For $e_{ij} \in \Delta^+_E$
    Randomly select $e_{kj'} \in T^+$, where $j' \notin E^i_t$
    Replace $e_{ij}$ in $G'_{t+1}$ with $e_{ij'}$
    Remove $e_{kj'}$ from $T^+$
For $e_{ij} \in \Delta^-_E$
    Randomly select $e_{kj'} \in T^-$, where $j' \in E^i_t$
    Add $e_{ij}$ to $G'_{t+1}$
    Remove $e_{ij'}$ from $G'_{t+1}$
    Remove $e_{kj'}$ from $T^-$
Return $(G'_{t+1})$
```

Table 3: Choice-based randomization method for assessing homophily

```
$Randomize^{inf}_{choice}(X_t, X_{t+1})$
$\Delta^+_X = \{x^v \in X_{t+1} \wedge x^v \notin X_t\}$ (added attributes in $t+1$)
$\Delta^-_X = \{x^v \in X_t \wedge x^v \notin X_{t+1}\}$ (dropped attributes in $t+1$)
// targets for random selections
$T^+ = \Delta^+_X$
$T^- = \Delta^-_X$
// randomize
For $x^v \in \Delta^+_X$
    Randomly select $x'^u \in T^+$, where $x^u \notin X^u_t$
    Replace $x^v$ in $X'_{t+1}$ with $x'^v$
    Remove $x'^u$ from $T^+$
For $x^v \in \Delta^-_X$
    Randomly select $x'^u \in T^-$, where $x^u \in X^u_t$
    Add $x^v$ to $X'_{t+1}$
    Remove $x^u$ from $X'_{t+1}$
    Remove $x'^u$ from $T^-$
Return $(X'_{t+1})$
```

Table 4: Choice-based randomization method for assessing influence

additions by $|e_{kj'} \in T : j' \notin E^i_t| - |e_{ij}|$, and similar values can be calculated for the deletions, as well as attribute cases. The assumption in the greedy approach is that nodes and attributes with many random options, at the start, are unlikely to run out of available options even if their assignments are decided later in the algorithm.

However, the greedy approach cannot guarantee that a node or attribute will have a valid random option at the point in the algorithm in which it is assigned. If this is the case, the particular node or attribute will retain its original assignment. This will not preserve the degree of nodes and attributes in the original data (since those original assignments may already have been given to others in the randomization). However, since the assignment for that node/attribute is identical to the original, it will prevent Type I errors from occurring. Ideally, after assigning additions and deletions to a node or attribute we could recompute the random options available to the remaining nodes/attributes and resort the list of edges/deletions. However, recomputing in this manner during the algorithm is computationally very expensive, and did not provide an increase in performance in practice, so we do not consider it further.

Finally, in addition to determining the significance of each effect, the sampling distributions for the randomization tests may also provide an estimate of the expected gain for each effect independently (where the full randomization is used to remove any joint effects from the estimation of gains due to homophily and influence):

$$\begin{aligned} E[gain_{hom}] &= \mu_{gains^I_R} - \mu_{gains^F_R} \\ E[gain_{inf}] &= \mu_{gains^H_R} - \mu_{gains^F_R} \\ E[gain_{oth}] &= \mu_{gains^F_R} \end{aligned}$$

We can define the overall expected gain as a function of these independent components to compare the proportion of gain due to each effect:

$$\begin{aligned} E[gain_{obs}] &= E[gain_{hom}] + E[gain_{inf}] + E[gain_{oth}] \\ &= [\mu_{gains^I_R} - \mu_{gains^F_R}] + \\ &\quad\ [\mu_{gains^H_R} - \mu_{gains^F_R}] + \mu_{gains^F_R} \\ &= \mu_{gains^I_R} + \mu_{gains^H_R} - \mu_{gains^F_R} \end{aligned}$$

5. SYNTHETIC DATA EXPERIMENTS

This section describes our investigation of the characteristics of the proposed randomization tests on semi-synthetic data. We are interested in two characteristics of statistical tests. (1) *Type I error*: the probability of rejecting a *true* null hypothesis (i.e., incorrectly concluding that there is a significant effect when there is not). (2) *Power*: the probability of rejecting a *false* null hypothesis (i.e., correctly concluding that there is a significant effect when there is one). If a statistical test has elevated levels of Type I error, that implies that many of the conclusions we draw from the test may be incorrect. In contrast, if a statistical test has low statistical power, that implies that legitimate performance differences may not be detected as significant.

We start with a base of real-world social network data and use the distributions of observed changes to generate data with different characteristics. On data with random changes we evaluate the Type I error of the test; on data with simulated homophily/influence effects, we evaluate the statistical power of the tests. The results show that our proposed tests have low Type I error and power increases as the number of data changes increase.

5.1 Data

Using a small subset of the real-world data gathered from Facebook (see section 6) as base for time t, we generated semi-synthetic data sets for time $t+1$ designed to either maximize or minimize the presence of homophily or influence effects. The synthetic data in time $t+1$ uses the distribution of changes in the original Facebook sample (i.e., number of adds/drops per person), however our data generation procedure chooses a new set of changes to ensure the data have certain characteristics (e.g., homophily).

More specifically, we hold G_t and X_t constant but generate new data for G_{t+1} and X_{t+1} to create three types of datasets:

Random: Changes are made randomly so there is no homophily or influence effect.

Homophily-rich: Attribute changes are made randomly, link changes are designed to maximize homophily.

Influence-rich: Link changes are made randomly, attribute changes are designed to maximize influence.

To generate data with homophily, link additions are chosen to maximize similarity among the incident nodes. When selecting a new link, the following weights are assigned to each possible link: $1+\gamma*(\#overlaps)$. Here *#overlaps* is the number of attribute values that the nodes share. The link weights are then normalized over all possible new links (pairs of unlinked nodes) to produce a probability of selecting any given link. The probabilities are used to randomly select the appropriate number of link additions, while weighting the likelihood heavily toward similar pairs of nodes. Dropping links is done in a similar manner except that this probability is calculated across current links in t and the inverse is used to drop links among pairs of nodes that are most dissimilar.

To generate data with influence, we add attribute values in a similar way. Here we again compute a weight for each attribute value: $1 + \gamma * (\#overlaps)$, but *#overlaps* is defined to be the number of neighbors who have the attribute value under consideration. Likewise the inverse is used to determine which attribute values are dropped.

5.2 Methodology

Type I errors correspond to cases when the null hypothesis is incorrectly rejected—in other words, false positive assessments of significance, when there is in fact no significant homophily/influence effect. To estimate the Type I error rate for each of the tests, we generated data with random changes in time $t+1$. Thus any observed gain in C is entirely random—so any assessment of significance will correspond to a Type I error. To evaluate the Type I error of the tests, we used the procedure outlined in Table 5.

Type II errors correspond to cases when the null hypothesis is incorrectly accepted—in other words, false negative assessments of significance, when there is in fact a significant homophily/influence effect. Power is the complement of the Type II error rate—the proportion of significant effects that are correctly identified $(1 - P(TypeII))$. To estimate Type II error of the tests, we generated data for time $t+1$ with either homophily or influence effects. Thus all observed gains in C should be deemed significant—any test that fails will correspond to a Type II error. To evaluate the statistical power of the tests, we used the procedure outlined in Table 6.

5.3 Experimental Results

We evaluated the Type I error rates of each test using semi-synthetic data, where the attribute and link changes are made at random. The power of the homophily and influence tests were evaluated using the Homophily-rich and Influence-rich data, respectively. For these experiments, unless otherwise noted, we used $N = 20, \alpha = 0.05$, $\gamma = 100$, and report average rates over 50 attribute values (i.e., group meberships).

The first column of Table 7 reports the Type I rates of the *choice-based* randomization test. Since we used $\alpha = 0.05$, we expect the error rates to be less than 0.05. This is indeed the case for both the homophily and the influence tests, indicating that the tests are likely to be accurate in practice.

```
$TypeIError(G_t, X_t, \Delta_G, \Delta_X, N, \alpha)$
For $i$ in 1..$N$
    Generate $G^i_{t+1}, X^i_{t+1}$ with random changes
    $numIncorrectSigTests = 0$
    For $j$ in 1..$N$
        $SignificanceTest(G_t, X_t, G^i_{t+1}, X^i_{t+1}, N, \alpha)$
        If significant: numIncorrectSigTests++
    $typeIError(i) = \frac{1}{N} numIncorrectSigTests$
$avgTypeIError = \frac{1}{N} \sum_i typeIError(i)$
```

Table 5: Method to measure Type I error rate.

```
$Power(G_t, X_t, \Delta_G, \Delta_X, N, \alpha)$
For $i$ in 1..$N$
    Generate $G^i_{t+1}, X^i_{t+1}$ with homophily/influence
    $numIncorrectSigTests = 0$
    For $j$ in 1..$N$
        $SignificanceTest(G_t, X_t, G^i_{t+1}, X^i_{t+1}, N, \alpha)$
        If not significant: numIncorrectSigTests++
    $typeIIError(i) = \frac{1}{N} numIncorrectSigTests$
$avgPower = \frac{1}{N} \sum_i (1 - typeIIError(i))$
```

Table 6: Method to measure statistical power.

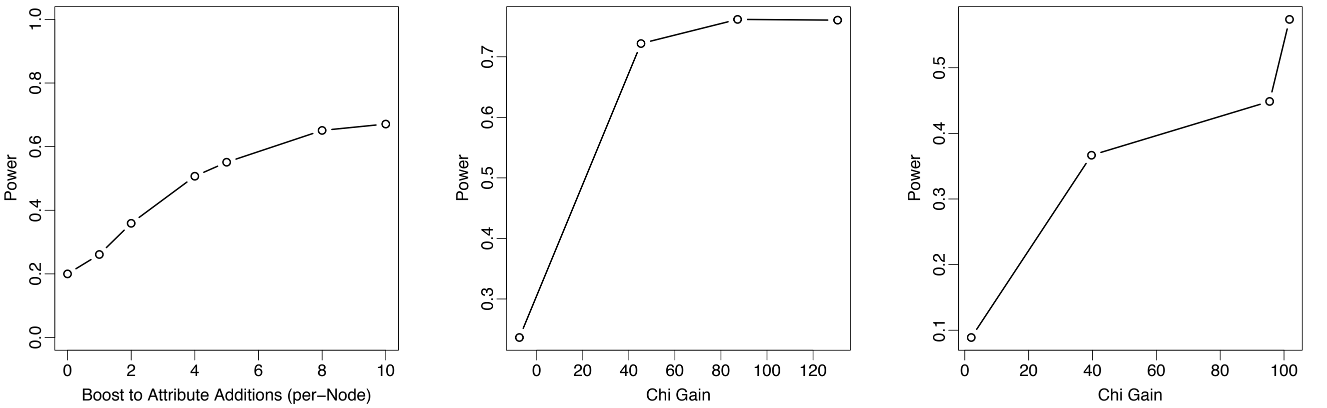
Next we evaluated statistical power. In this case we created data with only the effect we were trying to identify: (1) random attribute changes, homophily-based link changes (2) random link changes, influence-based attribute changes. These results are shown in columns 2 and 3 of Table 7.

The average power of the test for identifying homophily was 0.57, meaning an effect was correctly identified 57% of the time on data generated to include homophily. The average power for detecting influence based on our initial semi-synthetic datasets was significantly lower at 0.20. We conjectured that this was due to the presence of fewer changes in attribute values in the original data, compared to changes in the link structure. In particular, many fewer attribute values were being added—in the Facebook sample, the average number of link adds per node was 4.08 while the average number of attribute adds was 0.26 for the same set.

Figure 2(a) shows the interaction between the number of attribute additions and statistical power. As we increase the number of additions for each node in the graph, the power increase, reaching a maximum of 0.65 given a γ of 50. This is partially due to an increase in effect size (i.e., increase in overall level of influence) but the quantity of attribute changes adds an additional effect, over and above any increase in correlation gain due to increased influence.

The synthetic data generates homophily and influence by selecting link and attribute changes such that the correlation between node neighbors is maximized. If there are no such changes available, there will be less homophily and influence present in the synthetic data which degrades the power of the randomization tests. In addition, the distribution after randomizing will be closer to the original data as fewer changes can be randomized, which also reduces the power of the tests. In future work, we will attempt to tease apart these effects.

Figure 2(b)-2(c) shows the increase in statistical power



(a) Influence test power, as number of group additions increases.

(b) Influence test power, as effect size increases.

(c) Homophily test power, as effect size increases.

Figure 2: Power analysis of influence and homophily randomization tests.

	P(TypeI)	**Power**	**Power**
	Rand. Δ_X	Rand. Δ_X	Infl. Δ_X
	Rand. Δ_G	Hom. Δ_G	Rand. Δ_G
Homophily test	0.035	0.57	*na*
Influence test	0.04	*na*	0.76

Table 7: Type I error and power for the choice-based randomization method.

as we systematically decrease the effect size for either influence or homophily. Specifically, we varied γ to change the probability of selecting links and groups that produce autocorrelation in the synthetic data. For the influence test we used γ values of 1, 10, 20, 50 and for the homophily test values of 1, 10, 50, 100. We then plotted the expected power of the tests against the median chi gain of the groups. Greater chi gain places the group further from the null distribution, increasing the effect size. As expected, the power of each test decreases as the effect size decreases.

6. REAL DATA EXPERIMENTS

6.1 Data

We evaluated our approach on data from the public Purdue Facebook network. Facebook is a popular online social network site with over 250 million members worldwide. Members create and maintain a personal profile page, which contains information about their views, interests, and friends, and can be listed as private or public. Friendship links are undirected and are formed through an invitation by one user along with a confirmation by the other. To be affiliated with a University network, users must have a valid email account within the appropriate domain (e.g., purdue.edu), thus the members consist of students, faculty, staff, and alumni. The network we considered comprised more than 3 million public friendship links among 56,000 members. Users had an average and median degree of 46 and 81 respectively.

In addition to the friendship links, we considered a set of attributes corresponding to public *group membership.* Group membership information is posted in the users' profile pages. Each "group" maintains a separate page reflecting some interest (e.g., friends of AAAI), and users who share that interest can become members of the group.

		Influence	
		Significant	¬Significant
Homophily	Significant	7	111
	¬Significant	25	351

Table 8: Number of groups detected by randomization tests

For this work, we considered the set of 2648 (public) Facebook users belonging to the *class of 2011* student network. For the first time step, we used the friendship links and group memberships from March 2008. For the second time step we used friendship links and group memberships from March 2009. The students in this sample belong to 494 groups, so we consider each group membership as a binary attribute that can change from one time step to the next.

6.2 Experimental results

To investigate homophily and influence in Facebook, we computed the observed correlation gain for each group membership attribute from $t = 2008$ to $t + 1 = 2009$. We then applied the choice-based randomization procedure to determine if the gains exhibited significant homophily and/or influence effects. Due to the low type I error of the choice-based test, the discovered significant patterns are likely to be correct. However, the low power for influence identification (given the observed number of attribute changes in the sample) means that the influence test may not be able detect all effects in the data.

Of the 494 groups, notably 143 (29%) exhibited a significant correlation gain of some type. The assessments of significance are summarized in Table 8. Note that there are more groups that exhibit significant homophily effects (118) compared to significant influence effects (32). This is likely due the larger number of link changes, which results in higher power for the homophily test.

To explore the types of groups exhibiting each type of effect, we examined a set of group names selected at random from each cell in Table 8. Table 9 list some examples from each category.

Groups with significant homophily effects seem to include

Homophily and Influence (7 groups)
Purdue Habitat for Humanity
Tell 10 to Tell 10
Levee Tan
I started doing homework but I ended up on Facebook
Homophily Only (111 groups)
Purdue Capture the Flag
Honors Engineering Community 2007-2008
Boiler Gold Rush 2008
Purdue Opportunity Awards 2007-2008
Influence Only (25 groups)
NOBAMA IN 08
I bet I can still find 1,000,000 people who dislike George Bush
Hokay, so here's the Earth
i need numberss asap
No Effect (351 groups)
I support Welcome Home
We miss Cody Lehe
4-H alumni
Harrison Band

Table 9: Example groups with each possible combination of effect.

opportunities for members to meet in person. For example, *Boiler Gold Rush* is a freshman orientation program for Purdue where members are likely to meet, members of the *Capture the Flag* group presumably meet to play the game, and *Levee Tan* is a local tanning salon.

Groups with significant influence effects seem to have a political or activist aspect to them. This includes anti-Obama and anti-Bush groups as well groups like *Habitat for Humanity* and *Tell 10 to Tell 10*, which is a breast cancer awareness group.

Another group of note is *i need numberss [sic] asap*. This group was created by a user who had lost his phone and wanted his friends to post their phone numbers to the group wall. The members of this group already had friendship links between them and were joining the group to share phone numbers with other friends which naturally produces a detectable influence effect.

7. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel randomization procedure to investigate the *causes* of observed autocorrelation in network data. The test focuses on distinguishing social influence effects from homophily effects, and enables the accurate assessment of whether the effects are statistically significant.

The advantage of the proposed *choice-based* method includes: (1) a model-free approach which makes a relatively small number of assumptions about the data, (2) the ability to assess both homophily and influence effects, and (3) low Type I error with reasonable levels of power that increase as the number of changes in the data increase.

We evaluated our proposed methods on semi-synthetic social network data, showing efficacy of the approach. We then applied the choice-based method to a real-world dataset to investigate the aspects of its observed autocorrelation. Our analysis of a public university Facebook network shows that autocorrelation in group memberships is due to significant influence and homophily effects. However, different groups exhibit different behavior, which indicates homophily and influence vary with respect to group properties. In future work, we plan to investigate this variability more deeply, using multiple time steps to control for the amount of change expected in a single time step and investigating the association of effects with other group properties (e.g., density, popularity).

In addition, herein we have only considered randomization procedure for first-order effects (i.e., dyad-level dependencies). Considering second-order effects such as *structural similarity* and *community* level change may help to decompose additional external effects that are not explicitly encoded in the data, but are implicit in the temporal-relational dynamics.

Although we have investigated the characteristics of our modeling approach on social network data, the methods are broadly applicable to relational and network domains that are changing over time. Relational data often record information about people (e.g., organizational structure, email transactions) or about artifacts created by people (e.g., citation networks, World Wide Web) so it is likely that social phenomena such as homophily and influence will contribute to autocorrelated observations in a wide array of relational domains.

Acknowledgments

This research is supported by DARPA and NSF under contract numbers NBCH1080005 and SES-0823313. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, or the U.S. Government.

8. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, 2008.

[2] L. Anselin. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publisher, The Netherlands, 1998.

[3] S. Aral, L. Muchnika, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.

[5] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 307–318, 1998.

[6] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168, 2008.

[7] P. Doreian. Network autocorrelation models: Problems and prospects. In *Spatial Statistics: Past, Present, and Future*, chapter Monograph 12, pages pp. 369–389. Ann Arbor Institute of Mathematical Geography, 1990.

[8] H. Eldardiry and J. Neville. A resampling technique for relational data graphs. In *Proceedings of the 2nd SNA Workshop, 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2008.

[9] F. Guo, S. Hanneke, W. Fu, and E. Xing. Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007.

[10] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593–598, 2004.

[11] D. Jensen, J. Neville, and M. Rattigan. Randomization tests for relational learning. Technical Report 03-05, Dept of Computer Science, University of Massachusetts Amherst, 2003.

[12] B. Karrer, E. Levina, and M. Newman. Robustness of community structure in networks. *Physical Review E*, 77:046119, 2008.

[13] P. Marsden and N. Friedkin. Network studies of social influence. *Sociological Methods and Research*, 22(1):127–151, 1993.

[14] W. Mason, F. Conrey, and E. Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, 11:3:279–300, 2007.

[15] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–445, 2001.

[16] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:5594:824–827, 2002.

[17] T. Mirer. *Economic Statistics and Econometrics*. Macmillan Publishing Co, New York, 1983.

[18] J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the Workshop on Statistical Relational Learning, 17th National Conference on Artificial Intelligence*, pages 42–49, 2000.

[19] E. Noreen. *Computer Intensive Methods for Testing Hypotheses*. Wiley, 1989.

[20] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p*) models for social networks. *Social Networks*, 29:192–215, 2007.

[21] U. Sharan and J. Neville. Temporal-relational classifiers for prediction in evolving domains. In *Proceedings of the 8th IEEE International Conference on Data Mining*, 2008.

[22] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 655–664, 2008.

[23] T. Snijders, C. Steglich, and M. Schweinberger. Modeling the co-evolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences*, pages 41–71, 2007.

[24] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 870–878, 2001.

APPENDIX

A. PROOF OF THEOREM 1

Proof. As defined in Section 2:

$$C(X_t, G_t) = \chi_t^2 = \frac{(ad - cb)^2 \cdot N}{(a+b)(c+d)(b+d)(a+c)}$$

where a, b, c, d, N are defined with respect to X_t and $P_{R(t)}$. Let $\hat{k}_r$ and $\hat{k}_u$ be the expected number of *agreements* among related and unrelated people respectively in time $t+1$. Similarly, let $\hat{m}_r$ and $\hat{m}_u$ be the expected number of *disagreements* among related and unrelated people respectively. The changes to the contingency table will be as follows:

	$X_t^i = X_t^j = x$	$\neg(X_t^i = X_t^j = x)$
$(v_i, v_j) \in P_{R(t)}$	$\hat{k}_r - \hat{m}_r$	$-\hat{k}_r + \hat{m}_r$
$(v_i, v_j) \notin P_{R(t)}$	$\hat{k}_u - \hat{m}_u$	$-\hat{k}_u + \hat{m}_u$

Since there is influence in the data, the probability of agreement is higher for related pairs and the probability of disagreement is higher for unrelated pairs. Thus $\hat{k}_r > \hat{k}_u$ and $\hat{m}_r < \hat{m}_u$.

Subsequently, the change to the a, d diagonal is positive:

$$\begin{aligned}\Delta_{ad} &= (\hat{k}_r - \hat{m}_r) + (-\hat{k}_u + \hat{m}_u) \\ &= (\hat{k}_r - \hat{k}_u) + (\hat{m}_u - \hat{m}_r) \\ &> 0\end{aligned}$$

And the change to the b, c diagonal is negative:

$$\begin{aligned}\Delta_{bc} &= (\hat{k}_u - \hat{m}_u) + (-\hat{k}_r + \hat{m}_r) \\ &= (\hat{k}_u - \hat{k}_r) + (\hat{m}_r - \hat{m}_u) \\ &< 0\end{aligned}$$

However, the net change to N and each marginal is 0 (since $k = m$, $\hat{k}_r + \hat{k}_u = k$, and $\hat{m}_r + \hat{m}_u = m$). Let $[ad]' = (a + \hat{k}_r - \hat{m}_r)(d + \hat{m}_u - \hat{k}_u)$. From above we know that $[ad]' > ad$. Similarly let $[cb]' = (c + \hat{k}_u - \hat{m}_u)(b + \hat{m}_r - \hat{k}_r)$. Again, from above we know that $[cb]' < cb$.

Thus, the autocorrelation will be higher due to influence:

$$\begin{aligned}C(X_{t+1}, G_t) &= \frac{([ad]' - [cb]')^2 \cdot N}{(a+b)(c+d)(b+d)(a+c)} \\ &> C(X_t, G_t)\end{aligned}$$

□