ABSTRACT

A crucial task in the analysis of on-line social-networking systems is to identify important people — those linked by strong social ties—within an individual’s network neighborhood. Here we investigate this question for a particular category of strong ties, those involving spouses or romantic partners. We organize our analysis around a basic question: given all the connections among a person’s friends, can you recognize his or her romantic partner from the network structure alone? Using data from a large sample of Facebook users, we find that this task can be accomplished with high accuracy, but doing so requires the development of a new measure of tie strength that we term ‘dispersion’ — the extent to which two people’s mutual friends are not themselves well-connected. The results offer methods for identifying types of structurally significant people in on-line applications, and suggest a potential expansion of existing theories of tie strength.

INTRODUCTION

In a social network, an individual’s network neighborhood—the set of people to whom he or she is linked — has been shown to have important consequences in a wide range of settings, including social support [12,24] and professional opportunities[5, 15]. As people use on-line social networks to manage increasingly rich aspects of their lives, the structures of their on-line network neighborhoods have come to reflect these functions, and the complexity that goes with them.

A person’s network neighbors, taken as a whole, encompass a profoundly diverse set of relationships — they typically include family members, co-workers, friends of long duration, distant acquaintances, potentially a spouse or romantic partner, and a variety of other categories. An important and very broad issue for the analysis of on-line social networks is to use features in the available data to recognize this variation across types of relationships. Methods to do this effectively can play an important role for many applications at the interface between an individual and the rest of the network —managing their on-line interactions [9], prioritizing content they see from friends [1], and organizing their neighborhood into conceptually coherent groups [23, 25].

Tie Strength.

Tie strength forms an important dimension along which to characterize a person’s links to their network neighbors. Tie strength informally refers to the ‘closeness’ of a friendship; it captures a spectrum that ranges from strong ties with close friends to weak ties with more distant acquaintances. An active line of research reaching back to foundational work in sociology has studied the relationship between the strengths of

ties and their structural role in the underlying social network [15]. Strong ties are typically ‘embedded’ in the network, surrounded by a large number of mutual friends [6,16], and often involving large amounts of shared time together [22] and extensive interaction [17]. Weak ties, in contrast, often involve few mutual friends and can serve as ‘bridges’ to diverse parts of the network, providing access to novel information [5,15].

A fundamental question connected to our understanding of strong ties is to identify the most important individuals in a person’s social network neighborhood using the underlying network structure. What are the defining structural signatures of a person’s strongest ties, and how do we recognize them? Techniques for this problem have potential importance both for organizing a person’s network neighborhood in on-line applications, and also for providing basic insights into the effect of close relationships on network structure more broadly.

Recent work has developed methods of analyzing and estimating tie strength in on-line domains, drawing on data from e-mail [19], phone calls [27], and social media [14]. The key structural feature used in these analyses is the notion of embeddedness — the number of mutual friends two people share [22], a quantity that typically increases with tie strength. Indeed, embeddedness has been so tightly associated with tie strength that it has remained largely an open question to determine whether there are other structural measures, distinct from embeddedness, that may be more appropriate for characterizing particular types of strong ties.

Romantic Relationships.

In this work we propose a new network-based characterization for intimate relationships, those involving spouses or romantic partners. Such relationships are important to study for several reasons.

From a substantive point of view, romantic relationships are of course singular types of social ties that play powerful roles in social processes over a person’s whole life course [4], from adolescence [2] to older age [7]. They also form an important aspect of the everyday practices and uses of social media [28].

And they are an important challenge from a methodological point of view; they are evidently among the very strongest ties, but it has not been clear whether standard structural theories based on embeddedness are sufficient to characterize them, or whether they possess singular structural properties of their own [11, 18].

Our central finding is that embeddedness is in fact a comparatively weak means of characterizing romantic relationships, and that an alternate network measure that we term dispersion is significantly more effective. Our measure of dispersion looks not just at the number of mutual friends of two people, but also at the network structure on these mutual friends; roughly, a link between two people has high dispersion when their mutual friends are not well connected to one another.

On a large random sample of Facebook users who have declared a relationship partner in their profile, we find that our dispersion measure has roughly twice the accuracy of embeddedness in identifying this partner from among the user’s full set of friends. Indeed, for married Facebook users, our measure of dispersion applied to the pure, unannotated network structure is more effective at identifying a user’s spouse than a complex classifier trained using machine learning on an array of interaction measures including messaging, commenting, profile-viewing, and co-presence at events and in photos. Further, using dispersion in conjunction with these interaction features produces significantly higher accuracy.

The main contributions of our work are thus the following.

• We propose a new network measure, dispersion, for estimating tie strength. Given the ubiquity of embeddedness in existing analyses of tie strength, the availability of this new measure broadens the range of tools available for reasoning about tie strength, and about mechanisms for tie strength classification in on-line domains.

• We provide a new substantive characterization of romantic relationships in terms of network structure, with potential consequences for our understanding of the effect that such relationships have on the underlying social network.

• Given this characterization, we examine its variation across different conditions and populations. We find, for example, that there are significant gender differences in the extent to which relationship partners are recognizable from network structure, and that relationships are more likely to persist when they score highly under our dispersion measure.

It is also important to delineate the scope of our results. Our approach to analyzing romantic partnerships in on-line social settings is through their effect on the network structure, and the ways in which such relationships can be recognized through their structural signatures. As such, there is potential for it to be combined with other perspectives on how these relationships are expressed on-line, and the conventions that develop around their on-line expression; a complete picture will necessarily involve a synthesis of all these perspectives.

DATA AND PROBLEM DESCRIPTION

We analyze romantic relationships in social networks using a dataset of randomly sampled Facebook users who declared a relationship partner in their profile; this includes users who listed their status as ‘married,’ ‘engaged,’ or ‘in a relationship’. To evaluate different structural theories on a common footing, we begin with a simply stated prediction task designed to capture the basic issues. We take a Facebook user with a declared relationship partner, and we hide the identity of this partner. Then we ask: given the user’s network neighborhood—the set of all friends and the links among them—how accurately can we identify the relationship partner using this structural information alone? Figure 1 gives an example of such a Facebook user’s network neighborhood [21], drawn so that the user is depicted at the center; such diagrams are the ‘input’ from which we wish to identify the user’s partner. By phrasing the question this starkly, we are able to assess the extent to which structural information on its own conveys information about the relationships of interest.

We note that our question has an important contingent nature: given that a user has declared a relationship partner, we want to understand how effectively we can find the partner. There are different questions that could be asked in a related vein— for example, inferring from a user’s network neighborhood whether he or she is in a relationship. We briefly discuss the connections among these questions in a subsequent section, but the problem of identifying partners is our main focus here. As data for our analyses, we principally use two collections of network neighborhoods from Facebook. The first consists of the network neighborhoods of approximately 1.3 million Facebook users, selected uniformly at random from among all users of age at least 20, with between 50 and 2000 friends, who list a spouse or relationship partner in their profile. These neighborhoods have an average of 291 nodes and 6652 links, for an overall dataset containing roughly 379 million nodes and 8.6 billion links.

We also employ a smaller dataset — a sample of approximately 73000 neighborhoods from this first collection, selected uniformly at random from among all neighborhoods with at most 25000 links. We refer to this sample as the primary dataset, and the larger dataset in the preceding paragraph as the extended dataset. We compute our main structural and interaction measures on both the primary and extended datasets, and these measures exhibit nearly identical performance on the two datasets. As we discuss further below, we evaluate additional network measures, as well as more complex combinations of measures based on machine learning algorithms, only on the primary dataset.

All Facebook data in these analyses was used anonymously, and all analysis was done in aggregate.

To evaluate approaches for our task of recognizing relationship partners from network structure, we start with a fundamental baseline — the standard characterization of a tie’s strength in terms of its embeddedness, the number of mutual friends shared by its endpoints [22]. Embeddedness has also served as the key definition in structural analyses for the special case of relationship partners, since it captures how much the two partners’ social circles ‘overlap’ [11, 18]. This suggests a natural predictor for identifying a user u’s partner: select the link from u of maximum embeddedness, and propose the other end v of this link as u’s partner.

We will evaluate this embeddedness-based predictor, and others, according to their performance: the fraction of instances on which they correctly identify the partner. Under this measure, embeddedness achieves a performance of 24.7% — which both provides evidence about the power of structural information for this task, but also offers a baseline that other approaches can potentially exceed.

Next, we show that it is possible to achieve more than twice the performance of this embeddedness baseline using our new network measure, dispersion. In addition to this relative improvement, the performance of our dispersion measure is very high in an absolute sense—for example, on married users in our sample, the friend who scores highest under this dispersion measure is the user’s spouse over 60% of the time. Since each user in our sample has at least 50 friends, this performance is more than 30 times higher than random guessing, which would produce a performance of at most 2%.

Theoretical Basis for Dispersion.

We motivate the dispersion measure by first highlighting a basic limitation of embeddedness as a predictor, drawing on the theory of social foci [10]. Many individuals have large clusters of friends corresponding to well-defined foci of interaction in their lives, such as their cluster of co-workers or the cluster of people with whom they attended college. Since many people within these clusters know each other, the clusters contain links of very high embeddedness, even though they do not necessarily correspond to particularly strong ties. In contrast, the links to a person’s relationship partner or other closest friends may have lower mbeddedness, but they will often involve mutual neighbors from several different foci, reflecting the fact that the social orbits of these close friends are not bounded within any one focus—consider, for example, a husband who knows several of his wife’s co-workers, family members, and former classmates, even though these people belong to different foci and do not know each other.

Thus, instead of embeddedness, we propose that the link between an individual u and his or her partner v should display a ‘dispersed’ structure: the mutual neighbors of u and v are not well-connected to one another, and hence u and v act jointly as the only intermediaries between these different parts of the network. (See Figure 2 for an illustration.)

We now formulate a sequence of definitions that captures this idea of dispersion. To begin, we take the subgraph Gu induced on u and all neighbors of u, and for a node v in Gu we define Cuv to be the set of common neighbors of u and v. To express the idea that pairs of nodes in Cuv should be far apart in Gu when we do not consider the two-step paths through u and v themselves, we define the absolute dispersion of the u-v link, disp(u, v), to be the sum of all pairwise distances between nodes in Cuv, as measured in Gu −{u, v}; that is…where dv is a distance function on the nodes of Cuv. The function dv need not be the standard graph-theoretic distance; different choices of dv will give rise to different measures of absolute dispersion. As we discuss in more detail below, among a large class of possible distance functions, we ultimately find the best performance when we define dv(s, t) to be the function equal to 1 when s and t are not directly linked and also have no common neighbors in Gu other than u and v, and equal to 0 otherwise. For the present discussion, we will use this distance function as the basis for our measures of dispersion; below we consider the effect of alternative distance functions. For example, in Figure 2, disp(u, h) = 4 under this definition and distance function, since there are four pairs of nodes in Cuh that are not directly linked and also

have no neighbors in common in Gu − {u, h}. In contrast, disp(u, b) = 1 in Figure 2, since a and e form the only pair of non-neighboring nodes in Cub that have no neighbors in common in Gu −{u, b}.