Social Networks 43 (2015) 39–47

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

The structure of online social networks mirrors those in the offline world

R.I.M. Dunbar[a,*], Valerio Arnaboldi[a,b], Marco Conti[b], Andrea Passarella[b]

[a] Department of Experimental Psychology, University of Oxford, Oxford, UK
[b] Institute of Informatics and Telematics of CNR, Pisa, Italy

ARTICLE INFO

Keywords:
Ego-centric social networks
Network scaling
Network structure
Facebook
Twitter

ABSTRACT

We use data on frequencies of bi-directional posts to define edges (or relationships) in two Facebook datasets and a Twitter dataset and use these to create ego-centric social networks. We explore the internal structure of these networks to determine whether they have the same kind of layered structure as has been found in offline face-to-face networks (which have a distinctively scaled structure with successively inclusive layers at 5, 15, 50 and 150 alters). The two Facebook datasets are best described by a four-layer structure and the Twitter dataset by a five-layer structure. The absolute sizes of these layers and the mean frequencies of contact with alters within each layer match very closely the observed values from offline networks. In addition, all three datasets reveal the existence of an innermost network layer at ~1.5 alters. Our analyses thus confirm the existence of the layered structure of ego-centric social networks with a very much larger sample (in total, >185,000 egos) than those previously used to describe them, as well as identifying the existence of an additional network layer whose existence was only hypothesised in offline social networks. In addition, our analyses indicate that online communities have very similar structural characteristics to offline face-to-face networks.

1. Introduction

The growth of digital communication (and, in particular, social networking sites) over the past decade has raised fundamental questions about the constraints that exist over both the size and the pattern of social relationships. In one sense, the implicit promise of the new technologies was that they would open up the vista of a social world that was intrinsically unlimited in size. This becomes of particular interest in the light of the finding that there appears to be a cognitive limit on the size of natural face-to-face social networks (Dunbar, 1993; Roberts et al., 2009; Sutcliffe et al., 2012). This limit is thought to arise out of a combination of a cognitive constraint and a time constraint.

The central cognitive constraint, known broadly as the social brain hypothesis, is based on the observation that, in primates, the typical size of social groups correlates closely with the size of the neocortex (Dunbar, 1992), and in particular with the more frontal units of the neocortex (Joffe and Dunbar, 1997, Dunbar, 2011). This seems to imply that in some way the information-processing capacity of the brain limits the number of relationships that individuals of a particular species can manage, thus limiting the size of groups because they become unstable and prone to fission when they exceed this size. Species with larger (frontal) neocortices manage to maintain coherence in larger groups than those with smaller neocortices. This proposal has since been given considerable support by evidence from a series of neuroimaging studies which have shown, for both humans (Lewis et al., 2011; Powell et al., 2012; Kanai et al., 2012) and monkeys (Sallet et al., 2013), that within-species variation in social network size correlates with the volumes of particular brain regions at the level of the individual. Powell et al. (2012) showed that, at least in humans, this relationship is mediated by mentalising competences. Mentalising competences (most commonly associated with theory of mind or mindreading, the ability to understand another individual's mental state) form a natural recursion running from first order (the state of self-consciousness) through second order (formal theory of mind) to fifth order in normal human adults, with a range in adults of around fourth to seventh order (Kinderman et al., 1998; Stiller and Dunbar, 2007). Powell et al. (2012) were able to show that there was a causal relationship in which the volume of the orbitofrontal cortex determined mentalising skills, and mentalising skills determined network size.

In addition, however, there is also evidence to suggest that time imposes a constraint. Time becomes important because it seems

* Corresponding author. Tel.: +44 1865 271314.
E-mail address: robin.dunbar@psy.ox.ac.uk (R.I.M. Dunbar).

http://dx.doi.org/10.1016/j.socnet.2015.04.005
0378-8733/Published by Elsevier B.V.

that the strength of a relationship is determined by how much time two individuals spend together. In humans, self-rated estimates of the emotional closeness for dyadic relationships (using a simple 0–10 analogue scale) correlate closely with the frequency of contact (Roberts and Dunbar, 2011; Arnaboldi et al., 2013a), and these in turn correlate with willingness to behave altruistically towards the alter in question (Curry et al., 2013). Similar findings have been reported for monkeys (Dunbar, 2012). One reason for this is that time (and maybe social/emotional capital) is limited (Miritiello et al., 2013) and individuals are forced to choose between investing their time and/or emotional capital thickly among a small number of alters or thinly among a larger number. Pollet et al. (2011a), for example, found that although extroverts typically had more individuals in their social networks than introverts, their average self-rated emotional closeness to these individuals was significantly lower. Similarly, Roberts and Dunbar (2011) found that individuals who had larger social networks distributed their available social capital (as indexed by their self-reported emotional closeness) more thinly than those who had smaller networks.

Three studies of digital datasets have sought to determine whether social networks online are also limited in size, and if so to what size. Pollet et al. (2011b) examined the offline social network of heavy and casual users of internet social networking sites, and found that they did not differ. Gonçalves et al. (2011) downloaded traffic among the followers of individual Twitter accounts and, using a criterion of reciprocated exchanges to identify meaningful relationships, concluded that Twitter communities typically averaged between 100 and 200 individuals. Similarly, in an analysis of email traffic among physicists, Haerter et al. (2012) found, using a similar definition to identify relationships, that there was a marked downturn in the rate at which additional members were acquired once communities exceeded 200 individuals.

Individuals do not, however, distribute their social effort (whether measured by time or by self-rated emotional closeness) evenly among the alters in their networks. Indeed, there is considerable evidence to show that, within natural social networks, individual alters can be ranked in order of declining investment by ego (e.g. Saramäki et al., 2014) and that these rankings fall into a natural series of layers with a scaling ratio of ~3 that yields breakpoints at around 5, 15, 50 and 150 alters (Zhou et al., 2005; Hamilton et al., 2007). These layers correspond to marked differences in both the frequency of contact with alters and in rated emotional closeness (Roberts et al., 2009; Sutcliffe et al., 2012), seemingly reflecting a combination of temporal and cognitive constraints that give rise to the layered structure of networks.

We here combine these two sets of findings and ask whether, given that internet-based communication might be expected to bypass at least some of the time constraints that limit face-to-face networks, online social networks nonetheless still exhibit the same kind of structuring. For these purposes, we examine three online datasets, two of them culled from Facebook (Viswanath et al., 2009; Wilson et al., 2012) and a Twitter dataset specially downloaded for the purpose (Arnaboldi et al., 2013b). In each case, we use specific algorithms to search for patterns in the data so as to determine, first, whether a layered structure similar to the one found in offline ego networks is present in the reciprocated traffic data collected from online environments, and then, if so, to identify the sizes of these layers.

2. Methods

2.1. Facebook dataset #1

Facebook dataset #1 was obtained before 2009 when the default privacy settings allowed users inside the same regional network to have full access to each others' personal data (Wilson et al., 2012). This dataset has been widely used for social network analysis (see for example Arnaboldi et al., 2012). The dataset covers the time span from the start of Facebook in September 2004 until April 2008 (it is publicly available for research and can be accessed at http://current.cs.ucsb.edu/facebook/, "Anonymous regional network A"). As explained in (Wilson et al., 2012), the dataset represents only a subsample of the original Facebook regional network, in terms of downloaded Facebook profiles (~56%) and their Facebook friendships (~37%). Although other analyses on Facebook ego network structure have been conducted using this dataset (Arnaboldi et al., 2012), here we will improve existing results through a more refined analysis of the dataset, obtaining more accurate results about the size and the composition of ego network layers.

The dataset was downloaded using a crawling agent that obtained the complete public profile information (including personal information and the list of Facebook friends), and the Facebook wall data of a set of users in a large regional network of Facebook. The agent followed the friendship links to obtain a large connected component of the regional network. The 44% of profiles in the regional network that was not been downloaded were profiles with restrictive privacy settings or users disconnected from the giant component. Despite the high number of missing profiles, some of their data is still present in the dataset. In fact, if a public profile of a user A was connected to a non-public profile B, the posts sent from B to A were still visible in A's Facebook wall. Moreover, B would appear in the friend list of A. Therefore, information exchanged on missing links *from non-public* profiles *to public* profiles is still available. We miss information related to posts (i) *from public* profiles (node A in our example) *to non-public* profiles (node B) and (ii) between non-public profiles. We discuss below how we estimate traffic related to (i). As for (ii), the amount of data collected for non-public profiles is usually lower than that of public profiles since their communication traces appear only indirectly inside the walls of other public users. For this reason, most private profiles appear as users with low Facebook usage, which we discard in our analysis. Given this, we argue that missing information about their mutual interaction is not particularly problematic for our purposes. Hence, we reasonably assume that, despite not containing all the possible communication records between users in the regional network, the dataset is still a valid representation of Facebook social network for the purpose of ego network analysis.

We managed to partly reconstruct missing information in respect of point (i) above, as follows. We cannot tell from the dataset itself which profiles are public and which are not because, for a given friendship relationship, the dataset only reports the number of (undirected) interactions (posts or photo comments) that occurred, and not the properties of the profiles of the users involved, or the detailed interaction log. Therefore, we do not know for which links in the dataset we are missing interactions in one of the two directions. The only information we have is the percentage of non-public profiles, i.e. 44%. For this reason, we have selected randomly 44% of nodes, and assumed that those are associated with the non-public profiles.[1] We have doubled the number of interactions on all the links of the ego networks of those nodes. This corresponds to assuming that these relationships are perfectly bi-directional, and the (unknown) amount of interaction from public to non-public profiles is the same as the (known) amount of interaction in the opposite direction. We can expect that this process makes our results for internal layers accurate and less precise for external layers, for the following reasons. First, it is known that bi-directionality becomes stronger and stronger as

[1] We assume that the amount of non-public nodes without any connection to public nodes is negligible.