The scale-free distribution of electronic communication and the "Gravitational Law of Social Interaction"

Jacob Goldenberg\textsuperscript{1} & Moshe Levy\textsuperscript{2}

\textsuperscript{1}The Hebrew University, Jerusalem 91905, Israel. \textsuperscript{2}The Hebrew University, Jerusalem 91905, Israel. These authors contributed equally to this work.

The intensity of electronic communications, measured by the number of emails and Facebook links, is found to decrease inversely with the physical distance.

Abstract

Email, cellular phones, and other forms of electronic communications provide extensive quantitative data on social interactions. These data have been employed lately to investigate the structure of social networks [1-3], and the dynamics of human mobility [4]. One important aspect of social interaction which has not yet been studied, and which may be a key to understanding the development of social networks, is the dependence of the interaction intensity on physical distance. We investigate this dependence by analyzing two sets of data: email communications, and the records of 100,000 Facebook users. We find that both email volume and the number of Facebook links decrease inversely with the physical distance between the users. This finding can be interpreted as a "gravitational law of social interaction" asserting that the probability of a social link between two people who live at a distance $r$ one from the other is proportional to $1/r^2$. 
Our social interactions play a large part in defining who we are. Other than satisfying a basic human need, they influence how we feel [5], the choices we make [6,7], what we buy [8-10], and how we invest [11]. In this study we examine the relation between social interaction and physical distance by analyzing two sets of data: links of members on the Facebook network, and email communications. In traditional forms of communications, physical distance typically affects the time, cost, and ease of communicating. Thus, data about traditional communication does not necessarily reflect the underlying distribution of social links: even if we have the same number of short-range links and long-range links, because of the higher cost and longer time required for long-distance communications we may communicate less with our long-range contacts. In contrast, for electronic communications physical distance is irrelevant: it is just as easy to email someone who is 2000 miles away as it is easy to email our next-door neighbour. Thus, electronic communications provide an "uncontaminated" way to investigate the underlying distribution of social links.

Facebook is one of the largest and fastest-growing electronic social networks in existence today. In Facebook, users register, create personal profiles, and manage real-life and Internet friendships with other users.¹ Third parties are often invited to develop different applications for the Facebook platform. “My Personality” is such an application, which allows users to share their personal data, including their zip code. By examining pairs of friends (i.e. pairs of linked users) in the database, we constructed the distribution of link distances. The distance between any two linked users was calculated by converting each user’s zip code to geographical longitude and latitude and

calculating the distance. We collected data of 100,000 Facebook users, and found 1,297 linked pairs.

The distribution of Facebook links as a function of physical distance is shown in Figure 1. Panel A describes the empirical density function as a histogram. The empirical distribution is in very good agreement with a scale-free power-law distribution. The solid line shows the best power-law fit to the empirical data, with an exponent of -1.03 and a standard error of 0.03. Thus, the empirical distribution is in agreement with Zipf’s (1949) Law, according to which density is proportional to $1/r$, where $r$ is the distance (of course, for any finite-sized system, we consider a truncated Zipf distribution). Panel B shows the empirical cumulative distribution and the best fit to the cumulative Zipf distribution, i.e. the best-fit logarithmic distribution (recall that the cumulative distribution for a truncated Zipf density function is the log function). While the density estimate provides a good picture for most distances, for large distances the number of observations is insufficient to obtain clear results using this method. Therefore, for large distances we also employ the rank-distance method, with findings presented in Panel C of Figure 1. In this method, we ranked the links from the one with the greatest distance (rank $n=1$) to the one with the lowest distance. The figure presents the distance $r(n)$ as a function of rank, on a semi-logarithmic scale. The linear fit is consistent with Zipf’s Law: The linear relationship $\log(r(n)) = A - Bn$ implies that the number of links exceeding distance $r$ is $n(r) = \frac{A}{B} - \frac{\log(r)}{B}$. If we denote the total number of links by $N$, the proportion of the distances greater than $r$ is $n(r)/N$, and therefore the cumulative probability function $F(r)$ is $1-n(r)/N$: $F(r) = 1 - \frac{A}{BN} + \frac{\log(r)}{BN}$. The probability density function is the derivative of $F(r)$ with respect to $r$, i.e.
\[ f(r) = \frac{1}{BNr} \]. Thus, a linear relationship between \( \log(n) \) and \( n \) implies Zipf’s Law.

The correlation we obtain in between \( \log(n) \) and \( n \) is \( R = -0.997 \). To examine whether the obtained Zipf Law is a specific property of the Facebook network, below we employ a similar methodology to investigate the distribution of emails by distance.

Email is the most widespread form of electronic communications, with more than a billion users across the globe\(^2\). As it is extremely difficult to obtain detailed data on email communications due to obvious privacy issues, little is known about the volume of email communications as a function of the geographical distance between correspondents. In this study we collected these data by asking subjects to report the locations of the recipients of their last 50 email messages, and their own city of residence (the complete questionnaire is provided in the supplementary material section). Overall, we collected data for 4,455 email messages.

The distribution of e-mail distances is reported in Figure 2. Panel A reports the density function. As the geographic location is given only at the city level of detail, the resolution of the email data is less detailed than the Facebook data. Out of the 4,455 messages, 1818 (41\%) where sent within the same city, yielding a distance measure of zero. Obviously, this low resolution limits our ability to characterize the density function, especially for short distances. As a result, Panel A of Figure 2 provides only a rough description of the density function, which appears consistent with Zipf’s Law, but is not conclusive. Panel B provides a more detailed picture by presenting the cumulative

---

\(^2\) An October 2007 report by the technology market research firm The Radicati Group (http://www.radicati.com/) estimates that there are 1.2 billion email users in 2007, and expects this number to rise to 1.6 billion by 2011.
distribution. The solid line describes the empirical cumulative distribution, while the dashed line shows the best logarithmic fit. The cumulative distribution is in good agreement with Zipf's Law. To complete the picture, Panel C describes the rank-distance relationship, with a correlation of \( R = -0.989 \).

It is striking that both email and Facebook communications depend on distance in a very similar way, and that this dependence is given by the simple Zipf Law, encountered in many other branches of science [12-18]. A similar dependence has been found for internet routers [19]. What does this regularity imply about the probability of two individuals, with a distance \( r \) between them, being socially linked? Let us begin by making the simplifying assumption of identical "representative" individuals homogeneously spread over two-dimensional space. This implies that the number of individuals at distance \( r \) from a given individual is proportional to \( 2\pi r \) (the circumference of a circle with radius \( r \) around this individual). Denote the probability of a social link between two specific individuals with a distance \( r \) between them by \( p(r) \), and the number of links of distance \( r \) that the individual will have by \( f(r) \). \( f(r) \) is the number of "neighbours" at distance \( r \) multiplied by the probability of a link given this distance, i.e. \( f(r)=2\pi r \cdot p(r) \). As we empirically find \( f(r)=c/r \), this implies that:

\[
p(r) = \frac{c}{2\pi} \frac{1}{r^2}. \tag{1}
\]

Eq.(1), derived under the assumption of identical individuals, may be considered analogous to the gravitational law: the probability of a social link (the force) between two individuals (bodies of mass) is proportional to \( 1/r^2 \). Relaxing the assumption of identical individuals, by allowing individuals to have different susceptibilities (or unconditional probabilities) of being linked, \( m_i \) and \( m_j \), this analogy can be further extended to:
\[ p(r) = \frac{Gm_j}{r^2}, \]  

where \( G \) is a constant given by \( \frac{c}{2\pi \langle m \rangle^2} \).

Several scholars have predicted that electronic communications for which physical distance is completely irrelevant will fundamentally transform our social structure creating a "borderless society" [20-22]. Our empirical findings show that even though physical distance is technically irrelevant for electronic communications, we use these communication paths primarily for short distances, because most of our social contacts are local. The "Gravitational Law of social interaction", provided by eq.(1) and its generalization eq.(2), are derived from the empirical observation of the number of links as a function of \( r, f(r) \). Understanding the origin of this law, and its derivation as the result of some underlying stochastic process of human mobility and social link formation, presents a challenge for future research.


We are very grateful to Shirly Bitansky, Assaf Hefler, Yan Keren, Tal Levy, Abigail Rakover, Greg Ravikovich, Rom Schrift, Rachel Vovnoboy, and Jacob Shapiro, for their assistance and many helpful comments. Special thanks to David Stillwell for sharing his unique Facebook data with us. This research has been financially supported by the Israel Science Foundation, and by the Davdison, K-mart and Zagagi Funds.

Correspondence and requests for materials should be addressed to M.L. (mslm@huji.ac.il).
Figure legends

**Figure 1:** The Distribution of Physical Distances of Facebook Contacts.

Panel A shows the density of the distribution of Facebook contacts as a function of distance. The empirical distribution (histogram) is in very good agreement with a power-law distribution. The best power-law fit yields a power of -1.03 with a standard error of 0.03. Thus, the distribution is consistent with Zipf’s Law (1949), where the density decreases as 1/r. Panel B shows the empirical cumulative distribution (solid) and best fit to Zipf (log function, dashed). While it is hard to directly estimate the density for very large distances, Panel C employs the rank-distance analysis to show that Zipf’s Law holds for large distances as well. The linear relation between rank and distance on a semi-logarithmic scale implies Zipf’s Law (see text). The straight line shows the best linear fit to the empirical data, with R=-0.997.

**Figure 2:** The Distribution of Email Distances

The geographic distance between correspondents was calculated for 4,455 email messages originating in the USA. Panel A shows the density of the distribution of email messages as a function of distance. The empirical distribution is given by the bars, and the Zipf distribution is given by the solid line. Panel B shows the empirical cumulative distribution (solid) and best fit to Zipf (log function, dashed). Panel C provides the relationship between rank, n, and the corresponding distance r(n). A linear relation between rank and distance on a semi-logarithmic scale implies Zipf’s Law (see text). The straight line shows the best linear fit to the empirical data, with R=-0.989.
Figure 1.
Figure 2.