# Supporting Information

## Peterson et al. 10.1073/pnas.1320578110

### SI Text

In this paper, we have two aims: (*i*) to use maximum entropy to identify a general functional form for situations involving cost-minus-benefit constraints, and (*ii*) to fit data over a broad range of contexts. For *i*, maximum entropy predicts an exponential of a digamma function, provided that $k_0$ is known; it also shows how to compute a full distribution if we are given a single quantity, $\langle w \rangle$. For *ii*, our aim is just to do simple curve-fitting of data, given the mathematical form from *i*. In this case, we have no microscopic model for $k_0$, so we know neither $k_0$ nor $\mu^*$. In this case, our objective is not to find the full distribution from $\langle w \rangle$; rather, for *ii*, we are given the full distribution function, and our objective is to find the best values of $k_0$ and $\mu^*$ that fit it.

Our fitting procedure is as follows. We use the maximum-likelihood estimation function mle function in MatLab, with probability distribution specified by Eq. **7**. Table 1 shows estimated $\mu^*$ and $k_0$ values, with 95% confidence intervals. We calculate goodness-of-fit $P$ values using the Monte Carlo simulation procedure (based on the Kolmogorov–Smirnov test) described in ref. 1. The 13 datasets shown here are above the $P = 0.05$ statistical significance threshold proposed in ref. 1. $P$ values shown in Table 1 are based on 1,000 simulations for each dataset.

We fit Eqs. **7** to 13 data sets:

- Project membership on the social coding Web site GitHub (downloaded from konect.uni-koblenz.de).
- Edits made by users of the English-language Wikipedia.
- Interactions between users of the Pretty Good Privacy secure data transfer algorithm (2).
- Words occurring immediately after one another in a Spanish book (downloaded from konect.uni-koblenz.de).
- Deaths resulting from terrorist attacks from February 1968 to June 2006 (3).
- Wall posts by users to their walls on the social networking Web site Facebook, from a 2009 crawl of New Orleans Facebook (4).
- Pairwise, physical protein–protein interactions (PPI) of proteins detected in small-scale PPI network data, in yeast (*Saccharomyces cerevisiae*), fruit flies (*Drosophila melanogaster*), and humans (*Homo sapiens*) (5).
- Replies between users of the social news Web site Digg (6).
- Friendships between users of the Petster social networking site Hamsterster (downloaded from konect.uni-koblenz.de).
- Occurrences of unique words in the novel *Moby Dick* (7).
- Class–class dependencies in the software libraries JUNG and javax (downloaded from konect.uni-koblenz.de).

An overlay of all fits and datasets is shown in Fig. 3. Individual parameters and fits are shown in Table 1 and Fig. 2. A scatterplot of our maximum-likelihood estimates of $\mu^*$ and $1/k_0$ is shown in Fig. S1.


1. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661–703.
2. Boguñá M, Pastor-Satorras R, Díaz-Guilera A, Arenas A (2004) Models of social networks based on social distance attachment. *Phys Rev E Stat Nonlin Soft Matter Phys* 70(5 Pt 2):056122.
3. Clauset A, Young M, Gleditsch K (2007) On the frequency of severe terrorist events. *J Conflict Resolut* 51(1):58–87.
4. Viswanath B, Mislove A, Cha M, Gummadi K (2009) On the evolution of user interaction in Facebook. *Proceedings of the Second ACM SIGCOMM Workshop on Social Networks*. Available at http://conferences.sigcomm.org/sigcomm/2009/workshops/wosn/. Accessed June 1, 2012.
5. Patil A, Nakai K, Nakamura H (2011) HitPredict: A database of quality assessed protein-protein interactions in nine species. *Nucleic Acids Res* 39(Database issue, suppl 1): D744–D749.
6. Choudhury MD, Sundaram H, John A, Seligmann DD (2009) Social synchrony: Predicting mimicry of user actions in online social media. *Proc IEEE CSE'09*, 10.1109/CSE.2009.439.
7. Newman M (2005) Power laws, Pareto distributions and Zipf's law. *Contemp Phys* 46(5): 323–351.

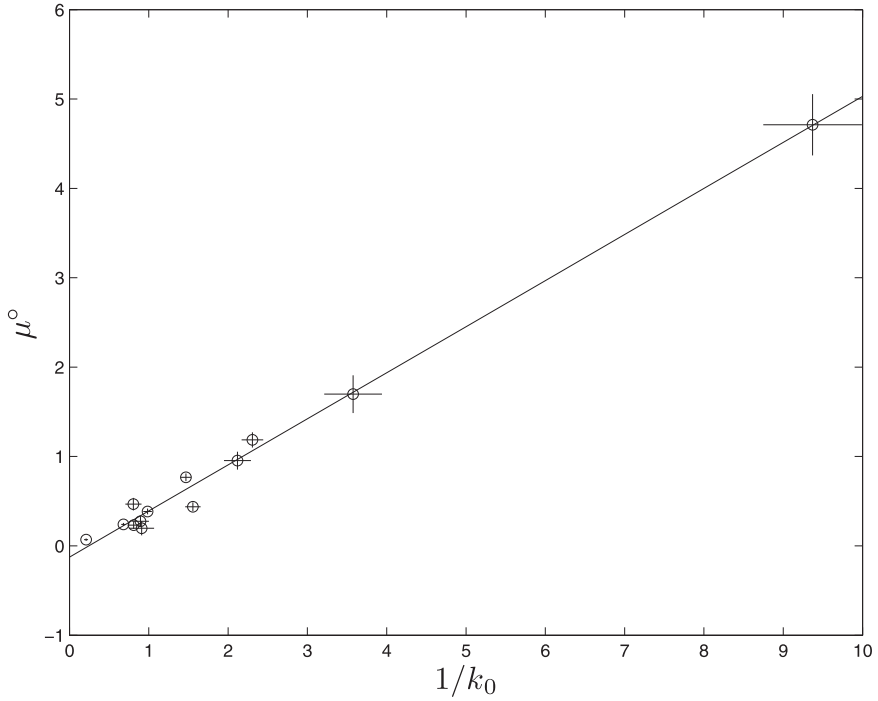


**Fig. S1.** The $\mu^{\circ}$ plotted against $1/k_0$ for the 13 data sets listed in Table 1. Error bars are 95% confidence intervals. Linear regression is shown as a solid line, $\mu^{\circ} = 0.516 k_0^{-1} - 0.125$ ($R^2 = 0.991$). Each point on this plot represents an empirical data set.