Available online at www.sciencedirect.com
SCIENCE DIRECT
Social Networks 28 (2006) 85–96
SOCIAL NETWORKS
www.elsevier.com/locate/socnet

The accuracy of small world chains in social networks

Peter D. Killworth [a,*], Christopher McCarty [b], H. Russell Bernard [c], Mark House [b]

[a] *National Oceanography Centre, Southampton, Empress Dock, Southampton SO14 3ZH, England*
[b] *Bureau of Economic and Business Research, University of Florida, Gainesville, FL 32611-7145, USA*
[c] *Department of Anthropology, University of Florida, Gainesville, FL 32610, USA*

Abstract

We analyse 10,920 shortest path connections between 105 members of an interviewing bureau, together with the equivalent conceptual, or 'small world' routes, which use individuals' selections of intermediaries. This permits the first study of the impact of accuracy within small world chains. The mean small world path length (3.23) is 40% longer than the mean of the actual shortest paths (2.30), showing that mistakes are prevalent. A Markov model with a probability of simply guessing an intermediary of 0.52 gives an excellent fit to the observations, suggesting that people make the wrong small world choice more than half the time.

Keywords: Small world chains; Social networks; Markov model

1. Introduction

The remarkable shortness of chains of social ties between any two individuals (the 'small world') continues both to surprise the public and to suggest multidisciplinary research with relevance to widely divergent areas, e.g. disease spread (Milgram, 1967; Killworth and Bernard, 1978b; Watts, 1999, 2004). Empirical studies involving envelope passing along chains of acquaintances (Travers and Milgram, 1969) demonstrated lower attrition than

* Corresponding author. Tel.: +44 23 80 596202; fax: +44 23 80 596204.
E-mail address: p.killworth@noc.soton.ac.uk (P.D. Killworth).

0378-8733/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2005.06.001

replications using the Internet (Dodds et al., 2003); both suggest a chain of order 6–7 links suffices to connect any two individuals. Proxy replications used scientific co-citation and organisational e-mail (Price, 1965; Newman, 2004; Adamic and Adar, 2005), and co-starring of film actors.[1] These proxies differ from the traditional small world in that there is no attrition of chains and, crucially, that individuals are not asked to make choices about the next step in a chain—indeed, researchers have full knowledge about all linkages available. The limited information available to individuals in the real world means that they make mistakes.

We know remarkably little about these mistakes. We do have good information about how individuals *make* small world choices, based on the target's location, occupation, hobbies and organisations, almost independent of the individual's culture (Killworth and Bernard, 1978a; Bernard et al., 1982, 1988). We do possess models of chain length for some ad hoc theoretical network structures (Newman, 2000; Watts, 2004), though not yet for estimates of the global distribution (McCarty et al., 2001; Hill and Dunbar, 2003; Killworth et al., submitted for publication) which appear to be long-tailed power-law-exponential. We have a suspicion that error occurs from the simple argument made by Pool and Kochen (1978), who connect two hermits in the U.S. by at worst a chain of seven via a storekeeper, a colleague of a storekeeper who knows a congressman, a congressman, and then repeating the chain in reverse to the other hermit. That chains appear empirically which are distinctly longer than seven must therefore be an indication of error. But we have no knowledge in general of whether individuals make the *right* choice to continue the chain, and what effect errors have on empirical results about the small world process. If, for example, error caused small world chains to appear longer than they actually are, conclusions about the spread of contagious diseases based on small world analyses would require heavy revision.

To study the effects of error, we need to know both the small world chains and the actual shortest paths possible. We address this issue by examining both the shortest possible paths, and the shortest paths selected by the small world approach, within a network about which we have full information.

2. The network

The network consists of $N = 105$ telephone survey interviewers, and, being work-defined, may yield results which do not necessarily hold for the global social network. A list of information about each network member, relevant to the work situation,[2] was created from an employee database. The list of members and information was presented to all members. Members indicated, for each person on the list, either that they knew the person, or provided a choice, from those they did know, for the link in a small world chain (together with one of seven work-related reasons for making that choice[3]). Six individuals provided no data (and are ignored in later computations which use $N' = 99$); there were instances of unintentional selection of self for a next link, treated as missing.

[1] The Internet Movie Database (http://us.imdb.com).
[2] Name, position within the survey, age, race, gender, time worked and schedule (shift pattern).
[3] The details provided, except for 'name'; a separate category of 'other' was also permitted.

Table 1
Means and correlations between quantities defined on network members

Quantity	pa−	nc+	nc−	*s*	*c*	*i*	*o*	ab	cb	Mean
pa+	**0.33**	**0.79**	0.03	**−0.34**	0.01	0.13	**0.38**	**0.40**	**0.23**	0.34
pa−	–	**0.29**	**0.65**	**0.24**	−0.03	−0.13	**−0.30**	**−0.21**	0.04	0.20
nc+	–	–	**0.42**	**−0.28**	0.11	0.10	**0.31**	**0.29**	**0.23**	0.62
nc−	–	–	–	−0.03	−0.03	0.02	0.02	0.00	0.16	0.55
s	–	–	–	–	0.18	**−0.66**	**−0.87**	**−0.82**	**−0.54**	0.87
c	–	–	–	–	–	−0.14	**−0.20**	**−0.25**	**−0.29**	0.58
i	–	–	–	–	–	–	**−0.42**	**0.54**	**0.54**	11.4
o	–	–	–	–	–	–	–	**0.85**	**0.40**	11.7
ab	–	–	–	–	–	–	–	–	**0.50**	0.01
cb	–	–	–	–	–	–	–	–	–	0.11

pa: path accuracy; + indicates inclusion of direct links, − indicates exclusion; nc: next choice accuracy (+, − as for pa); *s*: fraction of symmetric links; *c*: clustering coefficient; *i*: in-degree; *o*: out-degree; ab: actual betweenness; cb: conceptual betweenness. Betweenness is defined as follows: for member *k*, this is the average over all distinct connected pairs *i* and *j* of the fraction (number of paths from *i* to *j* passing through *k*/number of paths between *i* and *j*). Bold face indicates significance at the 5% level or better.

From these data, we extracted a 0–1 adjacency matrix **d** based on whether any link was known or not. This matrix is shown as a two-dimensional sociogram in Fig. 1a and b, with either interviewer type or length of time worked indicated by the colour scheme. More senior interviewers, measured either by position or months worked, tend to be more central, but otherwise there is little clear structure present, mostly because of shift patterns of working. (Curiously, there is a negative correlation, −0.22, evaluated over all pairs of individuals, between the number of common shifts they work and **d** itself.)

Fully 87% of the links in **d** are symmetric. The matrix satisfies the requirements for a small world network: its overall density $\sum_i \sum_{j \neq i} d_{ij}/(N'(N'-1))$ is low (0.12) but the mean density within those known to any individual (also known as the mean clustering coefficient; Watts, 2004) is high (0.58). The out-degree ($\sum_{j \neq i} d_{ij}$) distribution is long-tailed (the largest value being 76), qualitatively similar to estimates for the global network (McCarty et al., 2001). The in-degree ($\sum_{j \neq i} d_{ji}$) is qualitatively similar but with a much shorter tail (cf. Fig. 2). Summaries of these and other network measures are shown in Table 1. The majority of these are significantly correlated, suggesting that for these data at least, these network estimators may be measuring related information.

3. Small world paths

The shortest 'actual' path lengths through this matrix were calculated from the **d** matrix: of the 10,920 possible paths, 10,090 were found (the imperfect data causing omissions). The mean length was 2.30, S.D. 0.71. A choice matrix **c** was also created, whose (*i*, *j*)th entry lists the intermediary *i* chose to reach *j* (with value *j* if *i* listed *j* as known). Shortest 'conceptual' paths through **c** were also computed. Fully 2375 (21.7%) of these fail through reaching missing data. A further 2585 (23.7%) paths never terminate, reaching a cycling position (e.g. *i* chooses *j* chooses *i*). This could happen in a global small world problem,

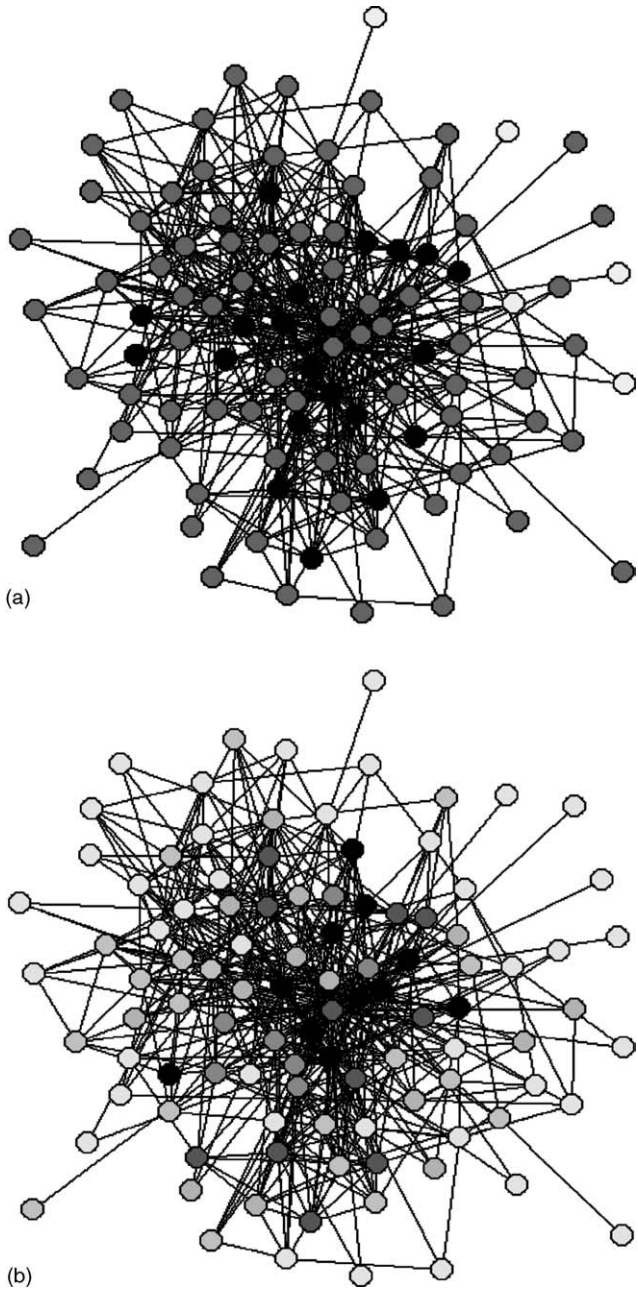


Fig. 1. Sociogram of the interviews. (a) Coloured by interviewer level (light grey means a newly hired interviewer, dark grey a general interviewer and black a senior interviewer). (b) Coloured by months worked in the laboratory, with darker colouring indicating a longer time worked.

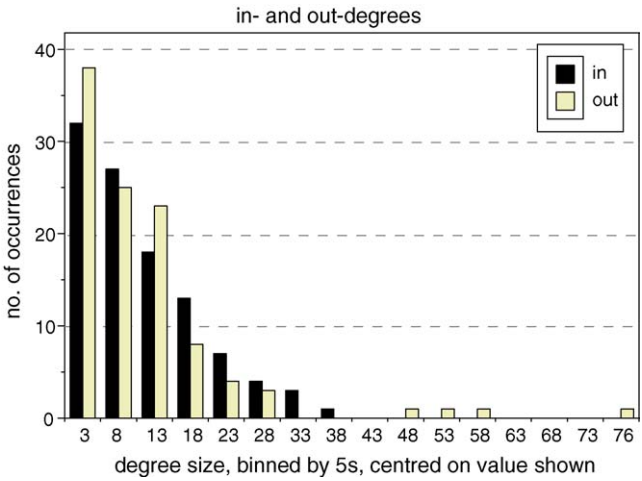


Fig. 2. Distribution of in- and out-degrees in the network, binned by 5s, centred on the value indicated (so '3' implies '1–5', etc.).

but in practice the chain would probably be lost. The remaining 5960 (54.6%) paths reach completion, with mean length 3.23, S.D. 2.06; this is 40% longer than the mean minimal path actually possible. Fig. 3a shows the distribution of path lengths.

Both actual and conceptual paths include the 1157 cases in which the respondent actually knows the target person (henceforth termed a 'direct link'). An alternative, and arguably more relevant, definition of mean path lengths ignores direct connections, yielding an actual mean length of 2.47, S.D. 0.57, against a mean conceptual length of 3.76, S.D. 1.94; the latter is now 50% longer than the equivalent actual mean length.

Actual paths (henceforth 'shortest' is assumed) are at most 5 in length, peaking at 2 and 3. Conceptual paths peak at similar lengths to the actual paths, but can be as long as 14. Error plays a large role in how the small world process occurs, even though individuals may choose from a fairly uniform average of three to four different possible choices of actual paths provided a conceptual path exists (Fig. 3b). This is in contrast to the roughly quadratic dependence of the mean number of different possible actual paths on the length of the actual path, which suggests a steadily more complex path structure as the path grows longer (Fig. 3c). Indeed, there are occurrences of very long actual chains. Fig. 3d shows a histogram of actual chain lengths across the $N'(N'-1)/2$ pairs of individuals; note the non-negligible contribution of paths longer than 21. Actual and conceptual path lengths are only weakly similar: 34% of the paths are the same length, but only 20% of conceptual chains longer than 1 are of the same length (Table 2).

The importance of each network member in actual and conceptual chains can be measured by the betweenness of the member (Freeman, 1980; Wasserman and Faust, 1994), defined in the caption to Table 1. For actual paths between two people there may be several routes involving one individual; for conceptual paths there can only be one (so that the mean conceptual betweenness, 0.107, is much larger than the actual, 0.013). 22% of the network

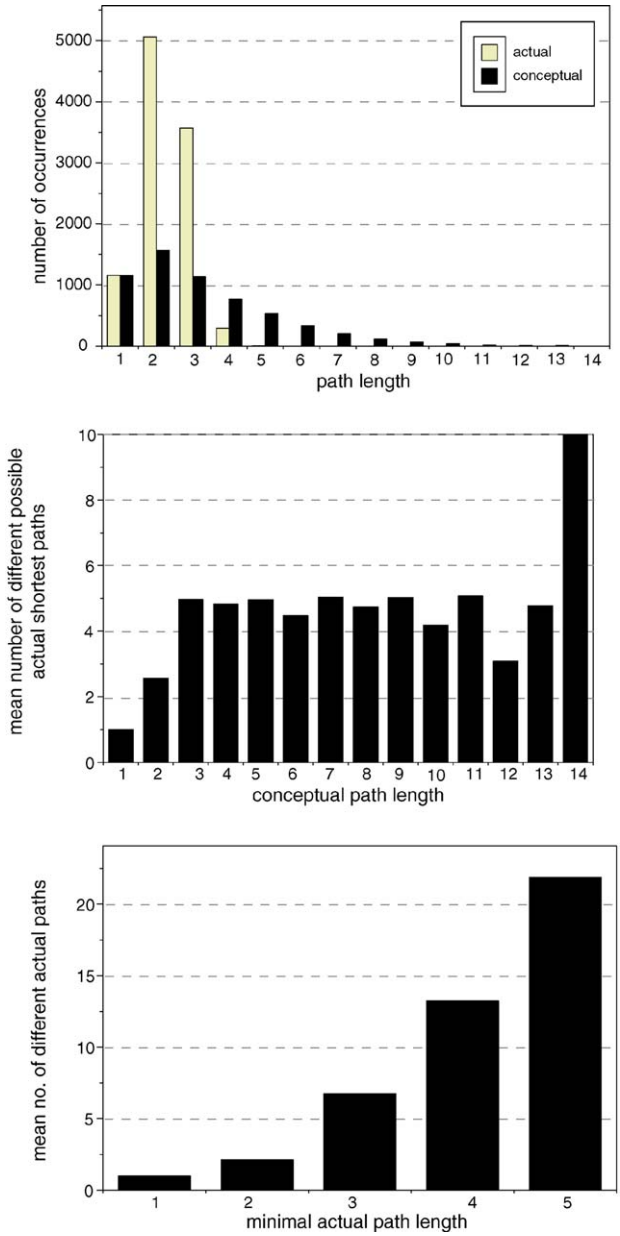
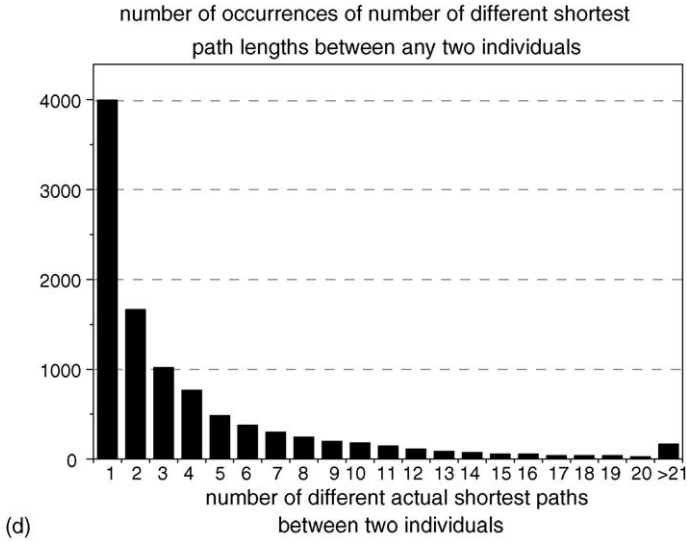


Fig. 3. Path lengths in the network. (a) Distribution of ‘actual’ and ‘conceptual’ path lengths. Actual paths are the shortest chains linked by acquaintanceship. Conceptual paths are chains linked by the choices made by members to reach each target person. (b) Mean number of different possible actual paths for a conceptual path of given length. (c) As (b), but for different minimal actual path lengths. (d) Number of actual shortest paths of various lengths across all pairs of individuals. (e) Mean length of actual and conceptual paths between pairs for each stated reason for conceptual choice.



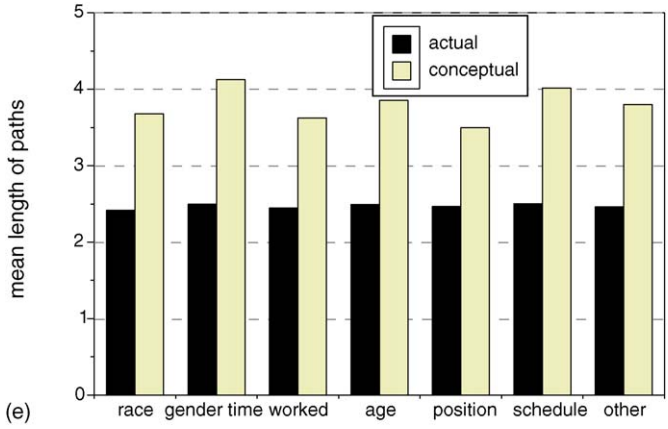


Fig. 3. (*Continued*)

Table 2
Row fractional distribution of conceptual path lengths

Conceptual length	1	2	3	4	5	6	7	8	9	10
Actual length										
1	1.0	0	0	0	0	0	0	0	0	0
2	–	0.47	0.21	0.12	0.08	0.05	0.03	0.01	0.01	0.0
3	–	–	0.31	0.23	0.17	0.11	0.07	0.05	0.03	0.02
4	–	–	–	0.42	0.23	0.12	0.13	0.04	0.04	0.02
5	–	–	–	–	0	0	0	1.0	0	0

For each actual path length, the fraction of conceptual paths of each length up to 10 is shown. The rows are normalized (i.e. non-terminating paths are ignored).

never appears in any conceptual path; only 2% never appear in actual paths. Actual and conceptual betweennesses are loosely related ($r = 0.50$). Members with larger betweenness tend to be those with larger in- and out-degrees (Table 1). Since we have no information on the strength of entries in **d**, we are unable to examine whether the hypothesis of the importance of weak links in small world chains holds (Granovetter, 1973).

The mean length of actual and conceptual paths varies significantly between the seven reasons given for conceptual choices (Fig. 3e), although the variation in the actual path lengths is numerically very small. The shortest conceptual paths are those chosen on the basis of the intermediary's position in the survey laboratory, suggesting that the limited hierarchical structure still provides effective routing. The accuracy of conceptual paths differs significantly between the seven reasons, with 32% of paths chosen by the intermediary's position being accurate (consistent with such being the shortest paths also) down to 17% accuracy for paths chosen on the basis of the race of the intermediary (White, 1970).

4. Accuracy of conceptual paths

We now discuss the accuracy involved in small world choices. There are two different accuracies to consider: *path* accuracy and *next choice* accuracy.

4.1. Path accuracy

A path is accurate if at all stages in the path a choice was made which was one of those in the list of existing actual paths, i.e. it has the same actual and conceptual lengths. Only 32% of conceptual paths were accurate. That figure includes direct links; if these cases are excluded, the fraction drops to 20%. In other words, in 80% of (non-trivial) small world chains, an error is made somewhere. Actual and conceptual path lengths are significantly correlated (0.14); the excess length of the conceptual over actual is significantly negatively correlated with the actual path length (−0.14). The equivalent figures without direct links are 0.37 and 0.13, respectively.

The path accuracy of each network member can be defined as the fraction of paths from that member which are accurate. The mean accuracy of network members is 0.34, decreasing to 0.20 if direct links are not included. Counterintuitively, path accuracy is significantly negatively correlated with the fraction of symmetric ties a member possesses (a symmetric tie might have involved knowing more about that network member). It is positively correlated with out-degree and both actual and conceptual betweennesses (all of these being well intercorrelated). Multiple correlation of individual path accuracy with the elements in Table 1 yielded $r^2 = 0.20^*$ (asterisk indicating 5% significance), though none of the best fit coefficients differed significantly from zero. The equivalent calculation without direct links yielded $r^2 = 0.13^*$, also with no significant best fit coefficients. (With these low variances we have made no effort to examine improvements to the fitting procedure.)

Thus, other features are involved in variations of path accuracy. Whatever these features are, they are not standard demographic factors: neither version of member path accuracy is

significantly correlated with the member's gender, race (black versus non-black, hispanic versus non-hispanic) or age. Neither time in the job nor type of interviewing job accounted for member path accuracy including direct links, though time in the job correlated significantly negatively with path accuracy without direct links. Since these factors comprised most of the reasons suggested by a pre-test, this again suggests a large amount of error in the small world process.

4.2. Next choice accuracy

The other definition of accuracy is less stringent: is the next choice in a chain one which actually moves the chain nearer to the target? (Operationally, if i chooses k as an intermediary to j, that choice is accurate if the conceptual distance from k to j is (one) less than that from i to j; recall that there may be several possible accurate choices.)

60% of all choices made were accurate, but these include the 1157 cases of direct links. Removing these reduces next choice accuracy to 48% (a figure which will be revisited in the next section). In other words, approximately half the non-trivial choices made by network members were incorrect.

The natural definition of the next choice accuracy of a network member is equivalent to that for path accuracy: the fraction of choices made by that member which are accurate. The mean value is 62%, dropping to 55% if direct links are removed. The multiple correlation of next choice accuracy (including direct links) with the elements of Table 1 yielded $r^2 = 0.17^*$, with a significantly positive coefficient for the member's in-degree. The equivalent without direct links produced only $r^2 = 0.04$, with no significant coefficients from any elements included.

Again, there was only one significant correlation between either version of member next choice accuracy with: gender, race, age, time in job or type of interviewing job. This sole significance was between next choice accuracy including direct links and gender of member (with increased accuracy for male network members).

4.3. A Markov model for path length

It is possible to account for the respondent errors which convert actual to conceptual path lengths by a simple Markov model, similar to models of the global small world problem (White, 1970; Hunter and Shotland, 1974; Killworth and Bernard, 1979). For a given network member and target i and j, the state of a chain can be classified by d_{ij} (here 1–5); chains which have reached the target are assigned a state of 0. To continue the chain we must specify the probability of transition to a new, or the same, state at the next stage. If respondents know the target, they are aware of this, so that chains at stage 1 reach the target at the next stage and terminate. For chains at state $n > 1$, we assume a constant probability α that the chain is broken (by the attrition possibilities above), and a second constant probability p that the respondent chooses correctly someone at distance $(n - 1)$. Otherwise, the respondent guesses, choosing someone at the same distance from the target with probability $(1 - p - \alpha)$. This specifies the transition probability matrix A_{ij} from state i to j in the next stage. In Hunter and Shotland's (1974) approach, the categories of the matrix were defined by the structural properties of the network (a university); in Killworth and Bernard's

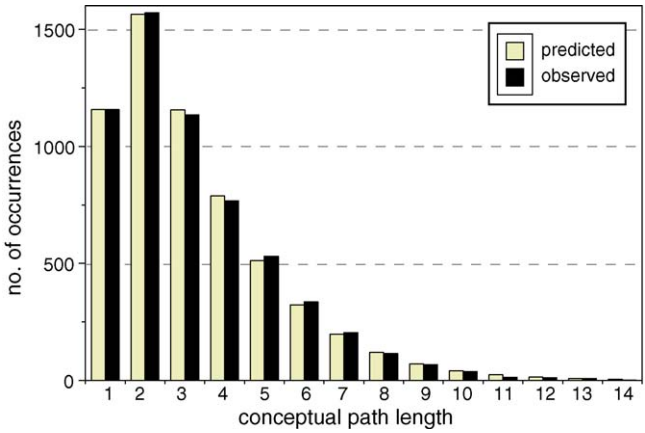


Fig. 4. Conceptual path lengths predicted from the Markov model, plus observed values ($\chi^2_{11} = 10.9$).

approach, the categories were defined by locational and occupational similarities with the target. There is no equivalent of either categorization for the work-related network here; hence, our reliance on the distance from the target.

The problem is initialised with a state vector $\mathbf{q}_0$ representing the entire network. We take this to be a zero entry for the target state, appended to the vector $\mathbf{M}$ of actual path lengths (Fig. 3a); $\mathbf{q}_0 = [0, \mathbf{M}]^{\mathrm{T}}$. The state of the chains then follows the rule $\mathbf{q}_n = A^{\mathrm{T}}\mathbf{q}_{n-1}$. Chains are followed for 14 steps to yield a predicted histogram of modelled conceptual path lengths, and the two free parameters α and p adjusted to minimise the discrepancy with observed values, yielding values $p = 0.308$ and $\alpha = 0.17$, making the probability of guessing 0.522. The results (Fig. 4) are indistinguishable from observations.

Thus, a simple model of state transition can be tuned to match observed values, implying that members of the network are inaccurate more than 52% of the time. This figure is consistent with that for mean next choice accuracy without direct links, which was 0.48 (cf. Section 4). This level of error is sufficient to increase mean path lengths 40% above the minimum possible; correcting for attrition would increase this still more (Hunter and Shotland, 1974; Dodds et al., 2003).

5. Conclusions

We have carried out the first investigation of accuracy in determining small world chains, by using a network in which the members provided information which enabled us to construct the entire actual and conceptual small world chains between all pairs of members. Consistently, a level of accuracy of around 50% is present (reminiscent of similar figures in other fields using human responses to questions; Bernard et al., 1984). This inaccuracy results in small world chains which are 40–50% longer than would be the case if ‘correct’ choices had been consistently made by network members.

These results suggest that inaccuracy in selection of small world chain intermediaries is predominant, with implications for deductions about (e.g.) infectious disease spread based on empirical small world research. This methodology should be extended to larger circumscribed networks to see if our findings hold. Of course, error in a closed system (as here) may be larger than in the global network: people may attempt to use the structure of the system in making choices, and misperceptions of that structure may result in error. In the global network individuals use target attributes as proxies for global structure (Bernard et al., 1982), as they are more aware that they cannot comprehend the structure, and so use only attributes of their acquaintances in an attempt to funnel into a local structure that increases their chances of completing the chain.

Acknowledgement

Sama Govinda collected much of the data at the University of Florida Bureau of Economic and Business Research.

References

Adamic, L., Adar, E., 2005. How to search a network. Social Networks 27, 187–203.

Bernard, H.R., Killworth, P.D., Evans, M.J., McCarty, C., Shelley, G.A., 1988. Studying social relations cross-culturally. Ethnology 27, 155–179.

Bernard, H.R., Killworth, P.D., Kronenfeld, D., Sailer, L.D., 1984. The problem of informant accuracy: the validity of retrospective data. Annual Review of Anthropology 13, 495–517.

Bernard, H.R., Killworth, P.D., McCarty, C., 1982. INDEX: an informant-defined experiment in social structures. Social Forces 61, 99–133.

Dodds, P.S., Muhamamed, R., Watts, D.J., 2003. An experimental study of search in global social networks. Science 301, 827–829.

Freeman, L.C., 1980. The gatekeeper, pair-dependency, and structural centrality. Quality and Quantity 14, 585–592.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Hill, R.A., Dunbar, R.I.M., 2003. Social network size in humans. Human Nature 14, 53–72.

Hunter, J.E., Shotland, R.L., 1974. Treating data collected by the “Small World” method as a Markov process. Social Forces 52, 321–332.

Killworth, P.D., Bernard, H.R., 1978a. The reverse small world experiment. Social Networks 1, 159–192.

Killworth, P.D., Bernard, H.R., 1978b. A review of the small-world literature. Connections 2 (1), 15–24.

Killworth, P.D., Bernard, H.R., 1979. A pseudomodel of the small-world problem. Social Forces 58, 477–505.

Killworth, P.D., McCarty, C., Johnsen, E.C., Bernard, H.R., Shelley, G.A. Investigating the variation of personal network size under unknown error conditions. Sociological Methodology and Research, submitted for publication.

McCarty, C., Killworth, P.D., Bernard, H.R., Johnsen, E.C., Shelley, G.A., 2001. Comparing two methods for estimating network size. Human Organization 60, 28–39.

Milgram, S., 1967. The small world problem. Psychology Today 22, 60–67.

Newman, M.E.J., 2000. Models of the small world. Journal of Statistical Physics 101, 819–841.

Newman, M.E.J., 2004. Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (Eds.), Complex Networks. Springer, Berlin, pp. 337–370.

Pool, I. de Sola, Kochen, M., 1978. Contacts and influence. Social Networks 1, 1–51.
Price, D.J. de S., 1965. Networks of scientific peers. Science 149, 510–515.
Travers, J., Milgram, S., 1969. An experimental study of the small world problem. Sociometry 32, 425–443.
Wasserman, S., Faust, K., 1994. Social Network Analysis. CUP, Cambridge.
Watts, D.J., 1999. Small Worlds. Princeton University Press.
Watts, D.J., 2004. The “new” science of networks. Annual Reviews of Sociology 30, 243–270.
White, H.C., 1970. Search parameters for the small world problem. Social Forces 49, 259–264.