Chapter 3

Fuzzy PFNETs: Coping with Variability in Proximity Data

Chris Esposito

Many of the quantities that cognitive scientists try to measure are more evasive than the physical constants measured by other sciences. The subjective similarity or distance measurements that are used to generate a Pathfinder network are no exception. As soon as one has decided to obtain similarity data from different individuals, one must also decide how to deal with any individual differences in these data. Several possibilities present themselves. If variability across subjects is high, it may be more useful to treat the set of subjects as several groups instead of one. Another approach is to treat each individual's data separately, generate individual solutions (Pathfinder networks, for example), and then compare the individual solutions in an attempt to synthesize a group solution. Another approach is to combine the individual datasets before the scaling method is applied; averaging the similarity ratings across subjects for each pair of items is a common step in preparing data prior to using Pathfinder. The issues of whether to use intersubject variability or not, and how to use it, must also be addressed.

The work to be presented in this chapter focuses on combining individual data matrices into a composite matrix and using the variability between the individual matrices in generating the network. We also introduce a new generation parameter, z, and discuss how it's value affects the generated network. To begin, we discuss some problems that arise from an interaction between the edge membership rule and how intersubject rating variability is used. Then we present some changes to the Pathfinder algorithms that add "missing" edges and often improve the statistical fit between network solutions and the data matrix they were derived from. Finally, some empirical evidence is presented to support the claim that the statistical version of Pathfinder produces networks that fit the original data better than networks generated by the nonstatistical version of the algorithm.

FUZZYPF—An Interval-Based Version of Pathfinder

When gathering subjective similarity data from several subjects, the standard procedure for getting a representative composite matrix is to compute the average of the corresponding elements in the individual subject matrices (Schvaneveldt, Durso, & Dearholt, 1985). Another standard procedure is that the average is the only composite measure that is used; intersubject variation for every pair is ignored. This has both advantages and disadvantages. One of the claimed advantages is that for any given pair of items, the average retains what is common among the individual ratings while filtering out relatively unimportant individual differences. If the variation across subjects for a given pair is fairly small, this advantage is probably real.

However, if there is substantial disagreement between subjects over the rating for a particular item pair, then the amount of variability increases sharply. This raises two

problems. First, given that the goal of the averaging is to get a representative composite value, it is no longer clear that a point value (rather than an interval) can genuinely represent widely disparate ratings. The second problem flows from how the edge membership rule compares path and edge lengths. The basic decision that occurs in generating a network is that if a path length is less than the edge length, the edge is *not* added. The problem is that *any difference*, no matter how small, is currently treated as if it were significant. As pointed out in Roske-Hofstrand and Paap (Chapter 4, this volume), many subjects will assign a slightly different score even to the same pair over time, so there appears to be an irreducible amount of variability in the score-assigning process. Failing to properly take this variability into account in network generation can result in edges being added to or omitted from the network based on differences between edge and path lengths that are sufficiently small as to not be significant. As previously mentioned, this can also reduce the fit between the network and the original data.

The proposed solution to these problems is to replace the point values for edge and path lengths with interval values constructed in such a way that both distance estimate intervals contain the "real" or "most likely" values for edge or path lengths. The basic decision then becomes one of comparing intervals to see if they overlap; if they do not and the path-length interval is below the edge-length interval, the difference between them is sufficiently large that it justifies not adding the edge. What follows below is a conceptual framework for developing a revised version of Pathfinder that uses intervals instead of point values in its decision making.

The basic model adopted is that for each edge (or pair of items) ij there is an independent random variable RV_{ij} and that the set of rating values RV_{ij} takes on across subjects constitutes a sample from the population of subjects of that type (e.g., the rating data for a command pair across a sample of UNIX experts can be used to estimate the "true" rating for that pair across the entire population of UNIX experts). The length of a path from node i to node j is also a random variable P_{ij} whose distribution is a function of the distributions of its constituent edge random variables.

As we remarked above, one view of using Pathfinder is that it consists of a sequence of decisions to be made, both by the user and the program. Several of these decisions are critically relevant here. First, how do individual ratings for a pair get combined into a composite? Second, how do individual edge lengths get combined into a path length? Third, how do we compare edge and path lengths? The statistical versions of these questions are given below, and it is those we shall attempt to answer.

(1) What is the distribution for each random variable?

(2) What is the appropriate measure of central tendency for the random variable? In other words, how do we compute its expected value?

(3) How do we determine the distribution for a path length given the distributions for its edge lengths?

(4) Given path and edge distributions, how do we compare them?

Before we begin to answer these questions it should be stated that the answers to questions 2–4 will be much easier to come by if we can demonstrate either that the random variables are each normally distributed, or that the normal distribution is a good approximation. The reason for this is that normal random variables are generally more well-behaved and easier to manipulate than those with other distributions. For example, Freund (1971) points out that normal random variables are closed under addition, minimum, and

maximum (the three basic operations in Pathfinder algorithms), while random variables with other distributions, such as exponential or geometric, are not.

Let us begin to answer these questions by taking another look at the basic operation of judging the similarity of two items. Without loss of generality, we can assume that these judgments are done on a scale of 1 to 10, with 1 meaning *not related* and 10 meaning *very related*. The ratings distribution across all pairs for any particular subject is generally bimodal, with one group of ratings at the high end and another group at the low end. This suggests that, at one level, the decision being made is simply *related* versus *unrelated*; the first outcome will be labeled *successes*, and the second outcome will be labeled *failures*. We can also divide the rating scale in such a way as to reflect this. Choose some position on the rating scale and let all scores less than the scale midpoint represent the *unrelated* outcome, with all scores greater than or equal to the midpoint as the *related* outcome. Without loss of generality, we can assume that for the moment the chosen position is the scale midpoint. Once the rating process has been recast in this way, a single rating of a single pair has all the characteristics of a Bernoulli trial: there are two possible outcomes, success (*related*) and failure (*unrelated*). If we repeat this procedure for a pair across n subjects then we have all the ingredients of a binomial distribution.

We are now in a position to answer the first question. As most statistics texts point out (for example, see Mendenhall, McClave, & Ramey, 1977), the Central Limit Theorem states that under certain conditions the normal distribution is a good approximation to the binomial distribution. The three parameters in a binomial distribution are n, p, and q, where p is the probability of success, $q = 1–p$ is the probability of failure, and n is the number of trials. Since we use one trial per subject, this is also the number of subjects. The two basic descriptors in a normal distribution are μ and σ, the mean and standard deviation. To determine whether the normal approximation will be adequate, calculate $\mu = np$ and $\sigma = \sqrt{npq}$. If the inequality, $0 \leq \mu \pm \sigma \leq n$, holds then the approximation will be reasonably good (Mendenhall, McClave, & Ramey, 1977). Assume, for a moment, that $p = q = 0.5$. We shall determine the minimum value of n for the normal approximation to be acceptable. Taking the left half of the inequality first, we have $0 \leq \mu - 2\sigma$. Substituting our values $\mu = 0.5n$ and $\sigma = \sqrt{0.25n}$ and simplifying, we get $0.4 \leq n$. So the left-hand inequality requires that the number of subjects is at least four. Taking the right half of the inequality next, we have $\mu + 2\sigma \leq n$. Substituting the same values for μ and σ and simplifying, we get $4 \leq n$. The right-hand inequality also requires that the number of subjects is at least four. For $p = 0.5$, as long as four or more subjects are used, the normal distribution is an acceptable approximation.

A closer examination of these bimodal distributions for a subjects ratings across pairs is likely to reveal that the region of the scale actually used for *related* judgments is much smaller than the entire top half. On our example scale that goes from 1 to 10, even if we require that *related* means a score of at least 9 (which leads to a p of 0.2), solving the above inequality yields a requirement that n be at least 16, which is still a very modest requirement.

Given that the assumption of a normal distribution is reasonable even for small numbers of subjects, we can now answer the other three questions posed earlier. If x is a normal random variable, then the appropriate measure of central tendency is the mean, or the average of the sample values. Therefore, averaging individual ratings for a pair to get a mean value is a defensible procedure. Since we are going to use sample means in calculating edge lengths and path lengths, the appropriate measure of variability is the equally familiar standard error of the sample mean, $\sigma_{\bar{x}}$, the formula for which can be found in any statistics text. As we shall see, an addition to the algorithm will be to calculate a matrix

containing pairwise standard errors of the mean in addition to the means we are already computing.

As defined in Schvaneveldt, Durso, and Dearholt (1987), the length of a path is the r^{th} root of the sum of the weights for the edges in the path, each raised to the r^{th} power, or the maximum of these edge weights when $r = \infty$. As pointed out earlier, the distribution of a sum of normally distributed random variables is also normal, as is the maximum and minimum. The mean (or expected value) of a sum of normal random variables is the sum of the means. Also, the variance of such a sum is the sum of the variances. Unfortunately, this result is exactly true only when the random variables in the sum are not raised to powers, that is, when $r = 1$ or $r = \infty$. The square of a normal random variable has a chi-square distribution, rather than a normal one (Freund, 1971). However, the Central Limit Theorem states that if a sample of n observations (with sample mean $\bar{x}$) is drawn from a population with an arbitrary distribution, finite mean μ and standard deviation σ, then as n increases the sample mean $\bar{x}$ will be increasingly normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. As a result, apart from some minor modifications to be described shortly, the path-length calculation does not change significantly.

At this point, let us review the decisions that have been made. Individual ratings are averaged to obtain composite edge lengths. Both edges and paths are represented by normally distributed random variables, with the latter as the sum of the former. The only issue left to resolve is how to compare an edge and a path and determine if one is shorter than another. The initial impetus for this revision of Pathfinder along statistical lines was that using point values to represent edge and path lengths led to some problems (e.g., the omission of edges), so we have replaced these point values by distributions. However, normal distributions extend infinitely in both directions, so a simple measure such as overlapping distributions goes too far in the opposite direction. The solution adopted in this work is to take the central portions of the distributions (the mean ± some user-specified amount on either side) and compare those intervals. This has the advantage of capturing the most likely values for edge and path lengths while being flexible enough to accommodate differing amounts of variability across different datasets and applications. An algorithm that incorporates all of these features is described below.

The data collection procedure for a single individual is unchanged. Given the set of individual distance matrices, we create two more matrices. The first is the average matrix W, where W_{ij} is the average of the corresponding entries in the subject matrices. The second is the variation matrix V, where V_{ij} is the value of one standard error of the mean for that pair and sample size.

FUZZYPF accepts these two data matrices as input. In addition to the q and r parameters, the user must specify the value of a third parameter z, which determines how much variability the program will use in making edge membership decisions. The parameter is called z because it determines how many standard errors the bounds on the edge weight intervals will be from their respective means, that is, their z scores. In order to create intervals of the desired size, the value zv_{ij} is added to and subtracted from w_{ij} in order to create the upper and lower bounds w_{iju} and w_{ijl} for that edge weight, so that as z increases, the intervals widen. The upper bound on the path length between nodes i and j (denoted P_{iju}) is the sum of the upper bounds on its constituent edge weights. Conversely, the lower bound on path length P_{ijl} is the sum of the lower bounds on its edge weights. The revised edge membership rule is that an edge is added to the network if the edge weight interval is less than or overlaps with the path-length interval. A necessary and sufficient condition for this to occur is $w_{ijl} \leq P_{iju}$.

A somewhat surprising result is how little the network generation algorithms need to be changed in order to accommodate the modifications just described. The one unfortunate result is that the amount of space required has gone up. Two additional $n \times n$ matrices are required, one to hold the lower bounds on the edge weights and one to hold the variation matrix. The matrix that is used to hold the current minimum path lengths starts out holding the upper bounds on edge lengths and is updated using these values as each algorithm proceeds. The most significant change in using Pathfinder is that the user must now choose a value for the new *z parameter*. The next section examines some of the issues surrounding this choice.

Choosing a z Value

The issues to be explored in this section deal with choosing a z value for the statistical version of Pathfinder that was just described. It should be fairly evident from the description of FUZZYPF that as z increases the intervals widen, and so it is increasingly likely that they will overlap and the edge will be added. As Figure 1 indicates, this results in increasingly dense networks as z increases. This raises the all too familiar question of what parameter value to choose. Some related questions are (1) how much improvement in network fit does FUZZYPF provide over nonstatistical versions, and (2) how does this improvement vary as a function of z?

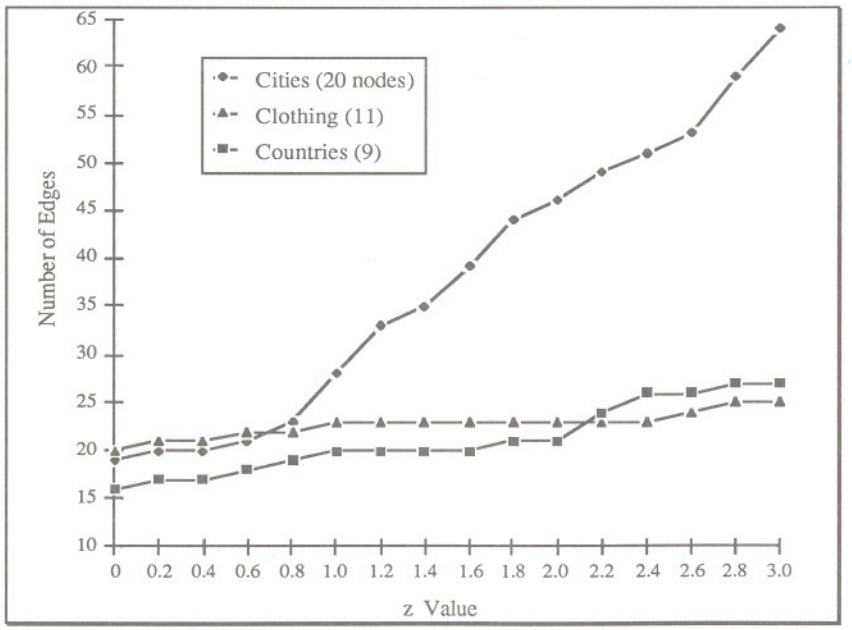

Figure 1. Number of edges as a function of z value. PFNETs($r = \infty$, $q = n$–1).

A useful way of thinking about these questions is to look at the set of edges not present in a network generated by a nonstatistical version of Pathfinder. These can be divided roughly into two types. Type 1 edges are those that clearly don't belong in the network because they are significantly longer than the shortest alternate paths. Type 2 edges are

those that are longer than the shortest alternate paths by an insignificant amount. Since the length differences are insignificant, the edge length and the path length should be considered as actually tied and the edge should be added, which often also improves the fit between the network solution and the original data. The goal is to find the z value that results in adding all of the Type 2 edges but none of the Type 1 edges.

The attempts to answer these questions are admittedly more exploratory than final. The approach used was to examine the fit between networks and original data matrices across several domains and a wide variety of z values. The datasets and measures of fit used here are also part of a larger study on the relationship between generation parameters, distance measures, and measures of fit between networks and the original distance matrices. For more details, see Esposito (Chapter 6, this volume). A brief summary of the relevant information is presented here.

The three domains used in this study were cities in New Mexico (20 terms, 12 subjects), items of clothing (11 terms, 20 subjects), and countries (9 terms, 9 subjects). For each domain, a set of networks was generated with $q = n–1$, $r = \infty$, and z varying from 0.0 to 3.0 in 0.1 increments. For each network we derived a distance matrix using the graph-theoretic definition of distance. Since the r value used in network generation entailed making only ordinal assumptions about the data, Spearman's ρ statistic was used to measure the fit between the derived and the original distance matrices (ρ also makes only ordinal assumptions).

Figure 2 presents a graph of fit as a function of z value for the three domains. Notice first that for this set of graphs, $r = \infty$ and $q = n–1$. The leftmost data point (network) is at $z = 0$. Since variability is effectively ignored at this z value, for each domain this network represents the PFNET($r = \infty$, $q = n–1$) as it would be calculated by any of the "regular" versions of Pathfinder and will serve as a reference standard to compare to networks generated with other values of z.

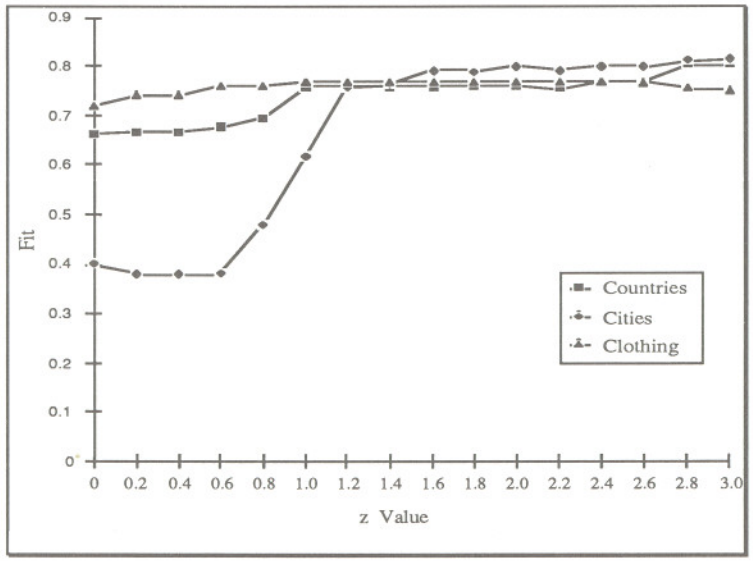

Figure 2. Network fit as a function of z value. PFNETs($r = \infty$, $q = n–1$).

As z increases, the edge-based fit curve for each dataset rises above where it was when $z = 0$, which strongly suggests that FUZZYPF produces networks that fit the original data better than networks produced by the regular version. In each of these datasets there is a fairly sharp rise in fit, followed by a leveling off, and then by a decline in fit as z gets rather large. The decline is easy to explain. As z increases, the networks get progressively denser. Since we are using a measure of fit that is based on the number of edges that separate two nodes, denser networks mean that everything is closer together; at some sufficiently large value of z, every pair of nodes will be nearly equidistant and so these distances will not fit very well with the original data matrix.

If we use the technique of looking for "elbows" in a fit curve, then all three of these curves have them, although different amounts of variability in the datasets means that they occur at different z scores. However, Table 1 shows that the same rising and leveling off in the number of edges added is also present. For the three domains used, the elbow in the edge curve (Figure 1) occurs at roughly the same z score that it occurs at in the fit curve (Figure 2). One can therefore make a case that a correspondence between the z value at which an elbow occurs in the fit curve and in the edges-added curve indicates that this is the z value required to create Type 2 edges in the network for that dataset and choice of q and r.

Table 1. Number of edges in various domains as a function of z value (values in parentheses are numbers of nodes).

	Domain				
z value	*bank(20)*	*fruit(11)*	*cities(20)*	*clothing(11)*	*countries(9)*
0.00	20	20	19	20	16
0.20	23	20	20	21	17
0.40	23	20	20	21	17
0.60	24	20	21	22	18
0.80	26	22	23	22	19
1.00	26	22	28	23	19
1.20	27	22	33	23	20
1.40	27	22	35	23	20
1.60	30	22	39	23	20
1.80	31	22	44	23	21
2.00	32	22	46	23	23
Complete	190	110	190	110	72
	Undirected	Directed	Undirected	Directed	Directed

Conclusions

In this chapter we presented a revised version of Pathfinder in order to deal with some statistical problems. The problems stem from the fact that there is often variability in subjective data, and in the old algorithm this variability was ignored for lack of a principled way of dealing with it. This often led to the omission of edges from the network because

they were slightly longer than the shortest alternate path, even though the difference in length was due to the random fluctuations in length rather than some statistically significant difference in length. The structure of the generated network was therefore affected in unpredictable ways that also often adversely affected the fit between the network scaling solution and the original data.

The new algorithm, FUZZYPF, explicitly incorporates a user-specified measure of variability by replacing the point values for edge and path length with intervals of user-specified width. The new criterion for edge membership in a Pathfinder network is that the edge is included if its length interval overlaps with the path-length interval. This new algorithm requires no more time than the old one, but does require two additional matrices, one to hold the standard error matrix and the other to hold edge weight lower bounds.

As a test of FUZZYPF, note that when $z = 0$, all variability is ignored and this version produces networks identical to the nonstatistical version. In order to compare the networks produced by FUZZYPF with those produced by a regular version of Pathfinder, we took five domains and computed the PFNET($r = \infty$, $q = n-1$) (with $z = 0$) for each one. We then let the z value range between 0.1 and 3.0 in increments of 0.1 and computed the fit (Spearman's ρ) between distance matrices derived from the networks and the original data. In every domain, the fit value rose and then fell as z increased, supporting the contention that for the "right" value of z, FUZZYPF produced better networks than the regular version of Pathfinder.

Despite the successful results reported above, it should not be concluded that the issue of how to deal with variability in subjective data when doing data scaling has been resolved in any final sense. Several related issues and approaches that were not addressed in this work are worth exploring. As mentioned at the beginning of this chapter, if between-subject variability is very high, a single network solution will not represent the group very well. On the other hand, if variability is very low, then using the averaged distance matrix by itself is probably sufficient. FUZZYPF offers the greatest advantage when the variability is moderate, so a more precise understanding of the size of this "moderate" range would be very useful. On a related note, different pairs of items will have differing amounts of variability in their similarity ratings, so some idea of what parts of the network are stable (less variable) and what parts are more variable would also be useful. A more general issue is the role of variability in data scaling in general. Some measure of how sensitive a clustering is to perturbations in the original distance data would be very useful. In addition, a version of multidimensional scaling based on interval rather than point values would likely produce some interesting results.