Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 1373–1386
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Predicting unobserved links in incompletely observed networks

David J. Marchette[a,*], Carey E. Priebe[b]

[a]*Naval Surface Warfare Center, Code Q21, Dahlgren, VA 22448, USA*
[b]*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA*

Received 9 January 2007; received in revised form 15 March 2007; accepted 16 March 2007
Available online 23 March 2007

Abstract

In this paper we consider networks in which the links (edges) are imperfectly observed. This may be a result of sampling, or it may be caused by actors (vertices) who are actively attempting to hide their links (edges). Thus the network is incompletely observed, and we wish to predict which of the possible unobserved links are actually present in the network. To this end, we apply a constrained random dot product graph (CRDPG) to rank the potential edges according to the probability (under the model) that they are in fact present. This model is then extended to utilize covariates measured on the actors, to improve the link prediction. The method is illustrated on a data set of alliances between nations, in which a subset of the links (alliances) is assumed unobserved for the purposes of illustration.
Published by Elsevier B.V.

Keywords: Random graphs; Social networks; Covert networks; Link prediction; Dot product graphs; Interstate alliances

1. Introduction

A graph is a pair $G = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ is a set of vertices (also called "actors") and E, the edge set (or "links"), is a set of unordered pairs of distinct vertices (this is a simple graph: we do not allow loops or multiple edges between vertices). In a directed graph (digraph) the edge set consists of ordered pairs, with (v, w) indicating an edge from vertex v to vertex w. In this paper we will consider only (undirected) graphs, but the ideas can be extended to the directed case. As is the convention, we will write $v_i v_j$ for $\{v_i, v_j\} \in E$.

The adjacency matrix $A = (a_{ij})$ is the $|V| \times |V|$ binary matrix with a 1 in position ij if and only if $v_i v_j \in E$. Since we are dealing with imperfectly observed graphs, where potential edges may not be perfectly observed, we encode an unknown edge with NA. This indicates that the link has neither been observed (1), nor is known to be missing (0).

Usually one assumes that an observed social network is fully observed: the actors and edges are complete and accurate. Typically the actors are a well defined small set of individuals or entities and the edge structure is easily obtainable through surveys or open sources. In this paper we consider the case in which the actors are actively trying to hide their links, or where the links are difficult to observe directly. The actors may have covariates associated with them which are assumed to be predictive, but not determinative, of the probability of the links between actors. We will assume that the actor set is known but the covariates on the actors may not always be perfectly observed. Unobserved covariates are also encoded as NA.

* Corresponding author.
E-mail addresses: dmarchette@gmail.com (D.J. Marchette), cep@jhu.edu (C.E. Priebe).

0167-9473/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.csda.2007.03.016

Given an imperfectly observed graph, one may need to expend resources to attempt to determine if a given link exists. For example, one may perform further surveys, contact the individuals directly, obtain a court order for a wire tap, or task a sensor to collect the relevant information. We would like to prioritize these tasks to ensure that the most likely links are examined first. This is the motivation for the approach we investigate here.

We will model the observed graph using the random dot product graph (RDPG) model described in Kraetzl et al. (2007), Scheinerman and Tucker (2007) and Marchette and Priebe (2007). This models the probability of an edge between two vertices according to the dot product of (d-dimensional) vectors assigned to the vertices. In Marchette and Priebe (2007) the model is constrained so that only a small number of distinct vectors are allowed, forcing the nodes to group according to their vectors, and reducing the complexity of the model.

The basic idea is to fit the RDPG model to the observed graph, then rank the potential links according to the probabilities induced by the estimated model. The hope is that the high probability links will turn out to be those that do in fact exist. This allows for an efficient assignment of resources to disambiguate the unknown relationships, and to accurately predict the graph from the imperfectly collected one.

To illustrate this approach, we investigate data representing alliances between a total of 173 nations collected from 1816 to 2000 (Gibler and Sarkees, 2004). The data are available at cow2.la.psu.edu, and we provide a processed data set (with more covariates than are considered in this paper) at www.ams.jhu.edu/∼marchette/igo.tgz. In this latter data set there are a total of 214 nations, but we will consider only the subset of these which have alliances between them. For each pair of nations, alliance is coded according to the type of alliance, but we will consider only the existence/absence of an alliance. There is a graph for each year, and we will treat these graphs independently in this paper, and will consider only a selected subset for the purposes of illustration. For some discussion of how to treat the time series of graphs, see Marchette and Priebe (2007).

2. Random dot product graphs

The most common random graph model in the literature is the Erdös–Renyí random graph (Bollobás, 2001), in which the edges are assumed to be independent. We do not believe that most interesting random graphs have independent edges, and so we seek a model that relaxes this requirement.

An RDPG is a random graph model (containing the Erdös–Renyí random graph as a sub-model) in which each vertex v_i is assigned a vector $x_i \in \mathbb{R}^d$. In this paper we will assume that $d = 2$, but in general part of the model selection task is to determine the appropriate value of d. The probability of an edge from v_i to v_j is a function of the dot product of the vectors:

$$P[v_i v_j \in E | x_i, x_j] := p_{ij} = f(x_i \cdot x_j). \tag{1}$$

In this paper we will set f to be a simple threshold:

$$f(x) = \begin{cases} 0, & x \leqslant 0, \\ x, & 0 \leqslant x \leqslant 1, \\ 1, & x \geqslant 1. \end{cases}$$

The vectors x_i are fixed, and new graphs are drawn from the collection of all graphs on n vertices according to the edge probabilities defined above. This model can be extended to directed graphs by assigning to each vertex two such vectors: an out-vector x^{O} and an in-vector x^{I} such that $p(i, j) = f(x_i^{\mathrm{O}} \cdot x_j^{\mathrm{I}})$.

One can think of the vectors x_i as vectors of covariates associated with the vertices, and one might wonder if these have any relationship to observed covariates. This is an interesting area for consideration, but we will take a slightly different approach to observed covariates. We consider the vectors x_i to correspond to latent variables, and seek to incorporate observed covariates within the model in addition to the x_i. Assume there are N measured covariates associated to each vertex, denoted α_i for the vector of covariates of vertex v_i. We will augment the probabilities in Eq. (1) as

$$p_{ij} = f(x_i \cdot x_j + \tilde{y}_i \cdot \tilde{y}_j), \tag{2}$$

where $\tilde{y}_i = y_i * \alpha_i$, and $*$ is component-wise multiplication. Thus the vector y_i corresponds to the weight associated with the covariate α_i associated to vertex v_i. In this model there are no interaction terms in the model for the covariates. Define $z_i = (x_i, y_i)$, the vector resulting from appending y_i to x_i. Then $\{z_i\}$ are the parameters of the model to be

fit. (In the model without covariates we set $z_i = x_i$.) The argument of f is a dot product, and we will write $z_i \cdot z_j$ for $x_i \cdot x_j + \tilde{y}_i \cdot \tilde{y}_j$.

It should be noted that in addition to being a generalization of the Erdös–Renyí random graph, the RDPG is a generalization of random intersection graphs (Karonski et al., 1999) and a sub-model of latent position models (Hoff et al., 2002, see also Hoff, 2005 for a very similar model).

Although computationally challenging, fitting the vectors to a given graph or set of graphs is relatively straightforward. Scheinerman (Kraetzl et al., 2007; Scheinerman, 2005; Scheinerman and Tucker, 2007) gives a linear algebra method that tries to minimize the Frobenious norm for the edge probabilities, and maximum likelihood is straightforward due to the fact that the probabilities are conditionally independent, given the vectors. Thus, the likelihood for the purely latent model is

$$L(z_1, \ldots, z_n) = \prod_{i \neq j} p_{ij}^{a_{ij}} (1 - p_{ij})^{1 - a_{ij}}, \tag{3}$$

with p_{ij} as above. (This is the likelihood in the directed graph case. In the case of an undirected graph, the product is taken over $i < j$. We will discuss below how to handle NA values in the adjacency matrix.)

Rather than allowing each vertex its own vector, we will assume a small number of distinct vertices, as is done in Marchette and Priebe (2007). For a graph $G = (V, E)$, a partition is a collection of subsets of vertices $P = \{P_1, \ldots, P_k\}$ such that $P_i \cap P_j = \emptyset$ for $i \neq j$ and $\cup P_i = V$. We will assume the number of partition cells K is known a priori. In a social network context, these groups might be club membership, interest groups, religious affiliation, or some unobserved grouping that one would like to discover. Thus, we are provided with a set of labels $\mathscr{L}$, and seek a map $h : V \to \mathscr{L}$. In the case where there are no vertex covariates, the likelihood is given by

$$L(x_1, \ldots, x_n; P) = \prod_{k,l=1}^{K} (f(x_k \cdot x_l))^{\tilde{a}_{kl}} (1 - f(x_k \cdot x_l))^{n_{kl} - \tilde{a}_{kl}}, \tag{4}$$

where $\tilde{a}_{kl}$ is the number of edges between vertices from partition cell P_k to partition cell P_l, and n_{kl} is the number of possible edges: $n_{kl} = |P_k||P_l|$ if $k \neq l$, $n_{kk} = |P_k|(|P_k| - 1)$. In the case of a graph rather than a directed graph we restrict the product to $k \leqslant l$ and $n_{kk} = |P_k|(|P_k| - 1)/2$.

This reduction does not quite work for graphs with vertex covariates, so in general we will use Eq. (3) with the z_i constrained to be constant on each partition. We will denote the likelihood in either case as $L(z_1, \ldots, z_n; P)$. This constrained random dot product graph will be denoted as CRDPG.

For fixed (known) values of K and d, a maximum likelihood estimate can be obtained via the following algorithm:

(1) Start with an initial set of vectors $\{z_j\}_{j=1}^{K}$ and an initial assignment of vectors to vertices (a partition P).
(2) For each vertex v_i (sequentially), select the partition cell P_j for which, after reassignment of the vertex to P_j, the likelihood is maximized:

$$P = \arg\max L(z_1, \ldots, z_n; P'),$$

where the arg max is taken over the K partitions P' differing from P by the assignment of v_i.
(3) With the assignment P fixed, select the vectors to maximize the likelihood:

$$(z_1, \ldots, z_n) = \arg\max_{(\hat{z}_1, \ldots, \hat{z}_n)} L(\hat{z}_1, \ldots, \hat{z}_n; P).$$

(4) Repeat from (2) until convergence.

At each stage of the algorithm, if there are missing values (NAs) in the adjacency matrix, we replace these with the estimated probability of the edge. The simplest form of this is to use the empirical estimate of the probability of an

edge between the given partitions. Thus if $v_i \in P_k$ and $v_j \in P_l$ we set

$$\hat{a}_{ij} = \frac{1}{|P_k||P_l|} \sum_{\substack{v \in P_k \\ w \in P_l}} I(vw \in E),$$

where I is the indicator function.

Note that the assignment in step (2) of the algorithm is sub-optimal: we only proceed through the vertices once sequentially, making the current assignment conditional on the previous assignments for the later vertices and the current assignments of the previous vertices in the list. This can be improved, presumably resulting in an algorithm that produces fewer iterations, at the expense of extra computation at this step. We will not consider these trade-offs here.

If instead of the computational strategy taken in (2) we tried all possible partitions (a finite but large number) then this approach would converge to the maximum likelihood estimate: in effect, we would select the configuration for which the likelihood is maximum. Instead we are checking a much smaller (greedily selected) subset of configurations, so the convergence is to a local maximum, not necessarily a global one. Since the number of partitions is finite, this converges.

Note that in the model without covariates, the least squares estimate for the x_i, assuming the partition is given, can be obtained immediately through the singular value decomposition: assume that each partition contains at least two elements. Define the reduced matrix M to be the $K \times K$ matrix where M_{ij} is the empirical probability of an edge from group i to group j. This empirical probability is computed using only the known edges/non-edges. Then if we write $M = UDV'$ using the singular value decomposition, the set of the first d eigenvectors in U (scaled by the singular values) is the optimal (in the least squares sense) choice for the $\{x_i\}$. If we are in a directed graph, the out-vectors come from U and the in-vectors from V.

There are several model selection problems to consider. First, one must decide on the dimensionality d of the vectors x_i. In a particular application, this might be in part driven by scientific considerations. One may hypothesize that there are two main factors defining the relationships, and thus choose $d = 2$ and observe the model fit. It should be noted that we want to pick d as small as possible, due to the additional variance in the estimators as the dimension increases. For the purpose of this work we will set $d = 2$, as was done in Marchette and Priebe (2007). Investigation of higher values of d showed no significant improvement in the models. We will also constrain $K = 2, 3$ as in Marchette and Priebe (2007). See that paper for more details about the model selection problem.

3. Results

The data correspond to 185 graphs, one per year. We will consider each year independently. Two of these graphs are depicted in Fig. 1. Fig. 2 shows the results of two CRDPG models fit to the data. The y-axis is the loglikelihood scaled by the number of possible edges in the graph. This allows the comparison of the models across time by removing the variability caused by the different number of non-isolated vertices. As can be seen, there is little difference between the $K = 2, 3$ models until the spike which occurs in 1936, after which the $K = 3$ model tends to do better. The vertical gray lines indicate the years upon which we will focus in this paper: 1866, 1886, 1906, 1936, 1941, 1950, 1970, 1977, 1990 and 1995. These years were chosen somewhat arbitrarily, but also focus both on areas where the two models agree (in likelihood) and where they disagree, as well as providing examples with various numbers of missing covariates (described below).

To evaluate the quality of link prediction in the CRDPG model, we use a methodology described in Hoff and Ward (2007):

(1) Split the set of potential edges into S disjoint sets (we will use $S = 4$ in this study; the edges are assumed to be missing at random).
(2) For each set, treat these edges as unknown by placing NA in the adjacency matrix.
(3) Estimate the CRDPG, and retain the probability associated with each missing (potential) edge.
(4) Rank the potential edges by their probabilities, and plot the number of tested edges against the proportion of true edges tested.

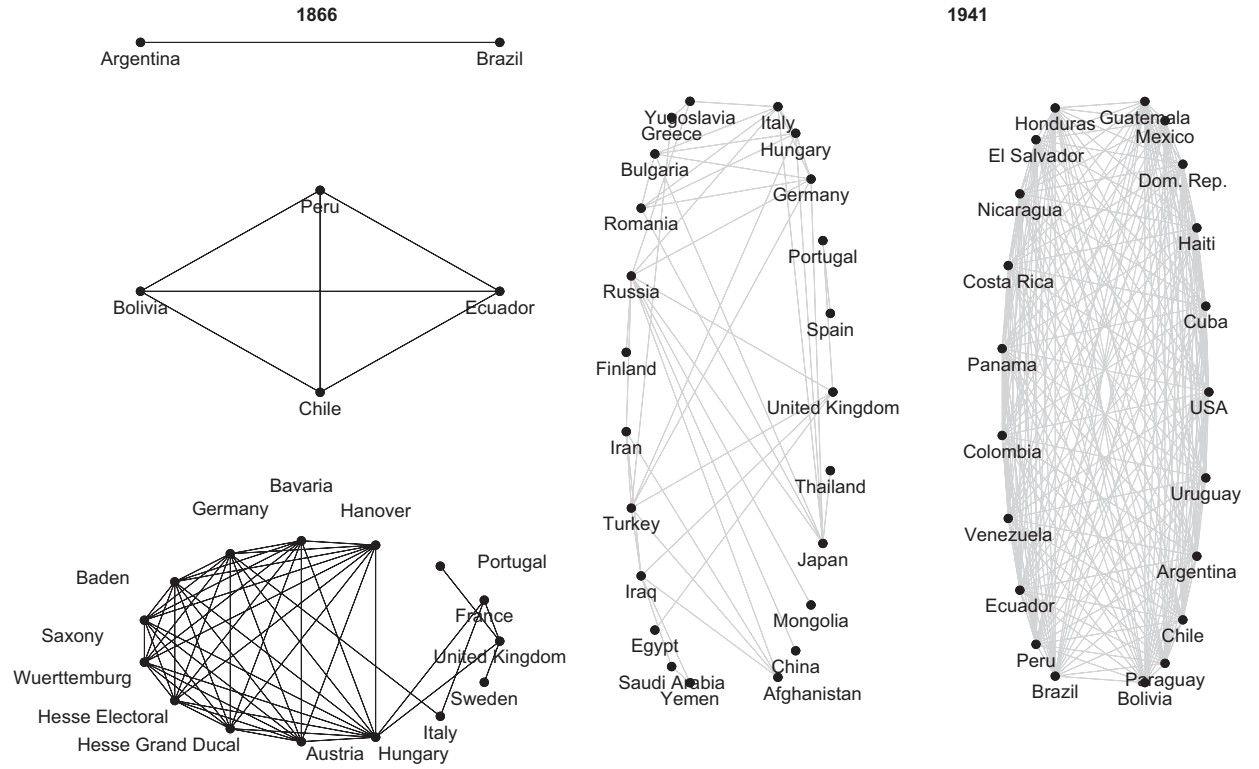

Fig. 1. Two of the alliance graphs, 1866 and 1941. The 1941 graph consists of two components: a complete graph on 21 vertices (the US and its Latin American allies) and a sparse graph on 23 vertices with 45 edges, corresponding to Europe and Asia.

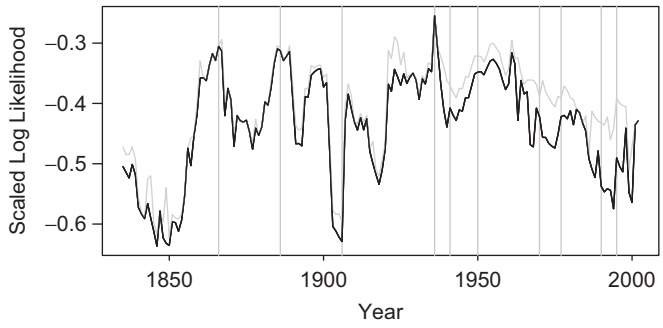

Fig. 2. Scaled loglikelihood values for $K = 2, 3$ across time. Each loglikelihood is scaled by the number of possible edges: $|V(t)|(|V(t)| - 1)/2$, where $V(t)$ is the set of non-isolated vertices at time t. The black curve corresponds to the $K = 2$ model; the gray curve corresponds to the $K = 3$ model. The vertical gray lines correspond to the years considered in the study.

This experiment is run 100 times (with different splits each time) to obtain an estimate of the confidence intervals for the predictions. In the fitting algorithm, the missing edges are imputed by the average of the probabilities for the corresponding groups, as discussed above. Fig. 3 shows one such run, for the year 1866.

We consider the two models, $K = 2$ and 3 with and without actor covariates. The covariates used are the amount of spending on the military, the amount of spending on energy, and a ± 10-point "democracy score". These covariates are first scaled as follows: for the first two, we transform the value α by taking $\log(\alpha + 1) + 1$. For the democracy score we shift so that the values range from 1 to 21. Finally, for each graph, each covariate is scaled so that it attains a maximum

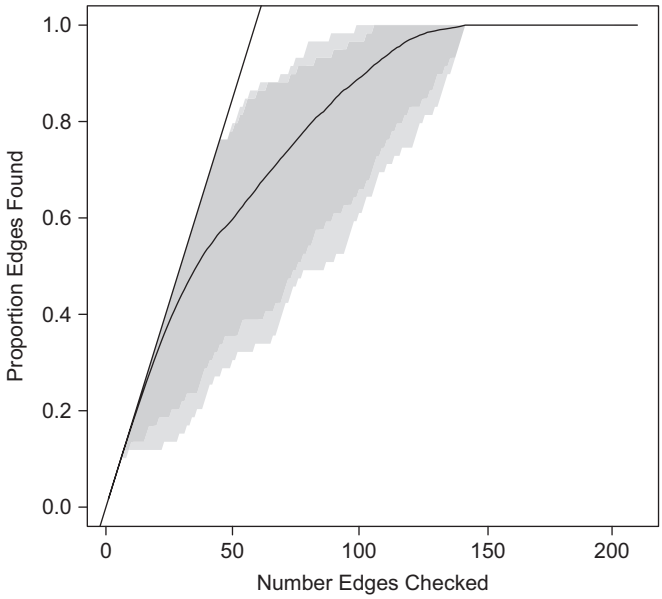

Fig. 3. An example of the results of 100 4-fold crossvalidation runs for the year 1866. The dark gray region is the empirical 90% confidence interval, and the light gray region shows the envelope of the runs. The mean is depicted as a solid curve. The $y = x$ line is shown for comparison purposes: this is the optimal curve obtained when each suspected link checked turns out to be a true link.

of 1. The plots in Figs. 4–7 show the results. Each year is depicted with two plots, one for each model. In each plot, the model using covariates is added as dashed/dotted curves indicating the mean and envelope for the model which incorporates covariates.

These plots show that, for the most part, about the first 40% or more of the links tested turn out to be true edges in the graph. This is a result of the tendency of these graphs to cluster. A graph that has two or three highly connected groups with relatively few edges in between would have the property that it would be relatively easy to predict missing edges, provided that one had identified the groups correctly. This seems to be the case for many of the graphs in the alliance data set. Note also that there is quite a difference in variability in the different years. This is associated with the amount of clustering: years consisting of relatively tight clusters (highly interconnected with few between-cluster ties) tend to have low variability in the crossvalidation, while loosely clustered or sparse graphs have high variability.

Note that the prediction performance does not completely agree with the likelihood on the question of model complexity. For example, in the years 1866 and 1886, where the two models ($K = 2, 3$ without covariates) have essentially the same likelihood (see Fig. 2), the more complex model produces a better prediction performance. This is because there is a trade-off between the reduction in variance provided by the "averaging" of information across the larger groups in the smaller model, and the reduction in bias provided by the better granularity of the larger model. The model selection performed in Marchette and Priebe (2007) used full edge information; it is not unreasonable that with missing edges a different level of complexity might be warranted.

Adding covariates to the model does not appear to improve the prediction, except possibly in 1886, $K = 2$. Even in this case, the improvement is not significant, and in that same year the $K = 3$ model shows a reduction of performance using covariates. One possible reason for this is that we have chosen covariates that are not well correlated with the attribute we are trying to predict. We did a small test to determine whether increasing the number of unknown edges would effect the performance, and the results are depicted in Fig. 8. Here, 50% of the potential edges are presented as unknown, and as can be seen there are several cases in which adding the covariates provides significant improvement in prediction. Thus we infer that the covariates are of some value in prediction.

There are missing values in the covariates (Table 1), and this was not taken into account in the above analysis. Thus, another potential reason for the lack of improvement in the 4-fold crossvalidation results is these missing covariate values. The number of missing covariate values is depicted in Fig. 9 and in Table 1. In the above model, missing data

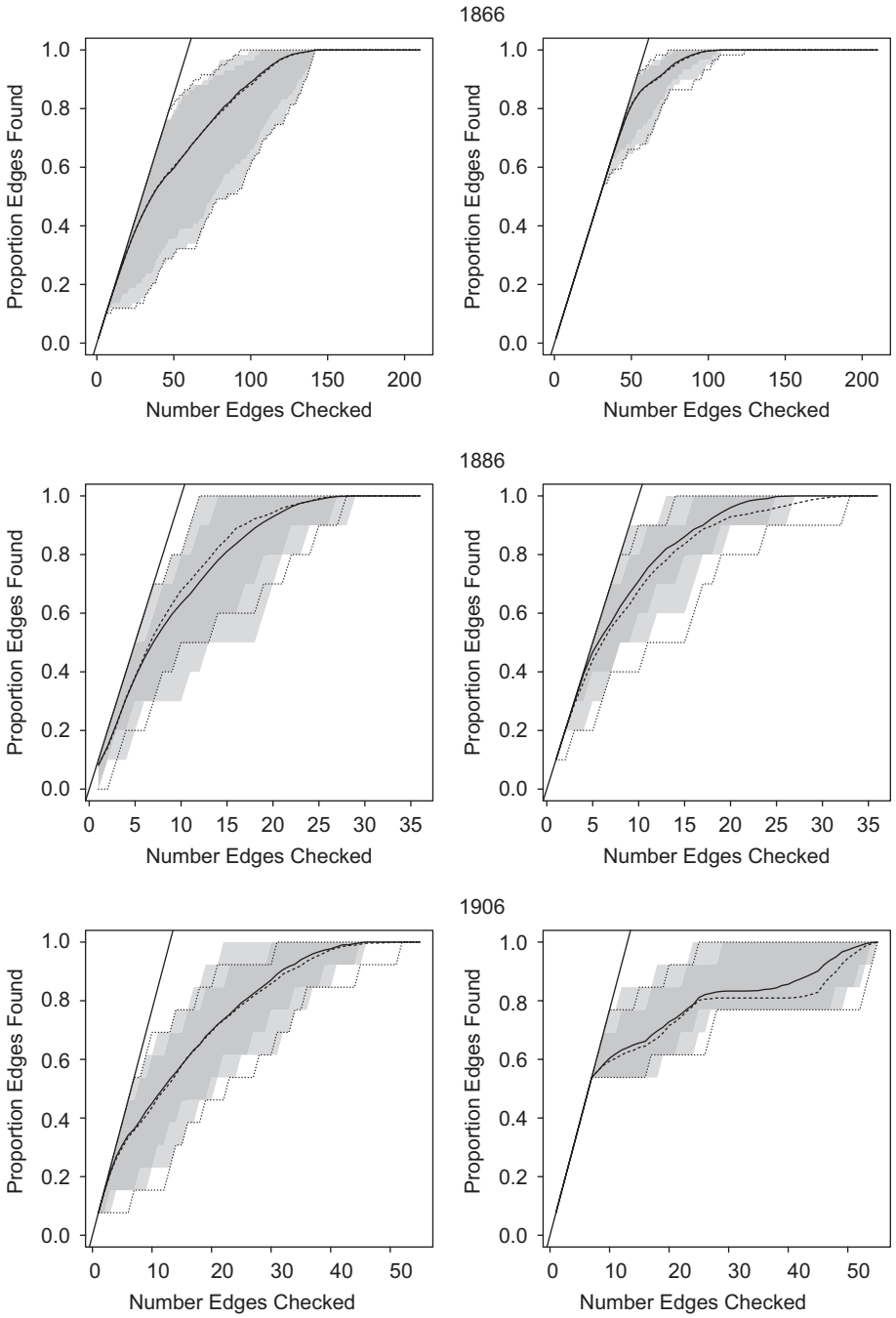

Fig. 4. Results for 100 4-fold crossvalidation runs for the years 1866, 1886 and 1906. These are paired by year, with the first plot corresponding to the $K = 2$ model, and the second plot corresponding to the $K = 3$ model. In each case, the results corresponding to the incorporation of the three covariates is overlayed as dashed curves indicating the means and dotted curves indicating the envelope.

were treated as 0 and thus did not effect the value of the probability obtained from the dot product (that is, we obtain the same value for the probability in the case of a missing value as we would if we did not utilize covariates in our model). We can try to mitigate this problem through a very simple imputation: if a value of a covariate is missing for a node in a given year, we replace the value with the last value for that covariate for that node in the past. The dotted curve

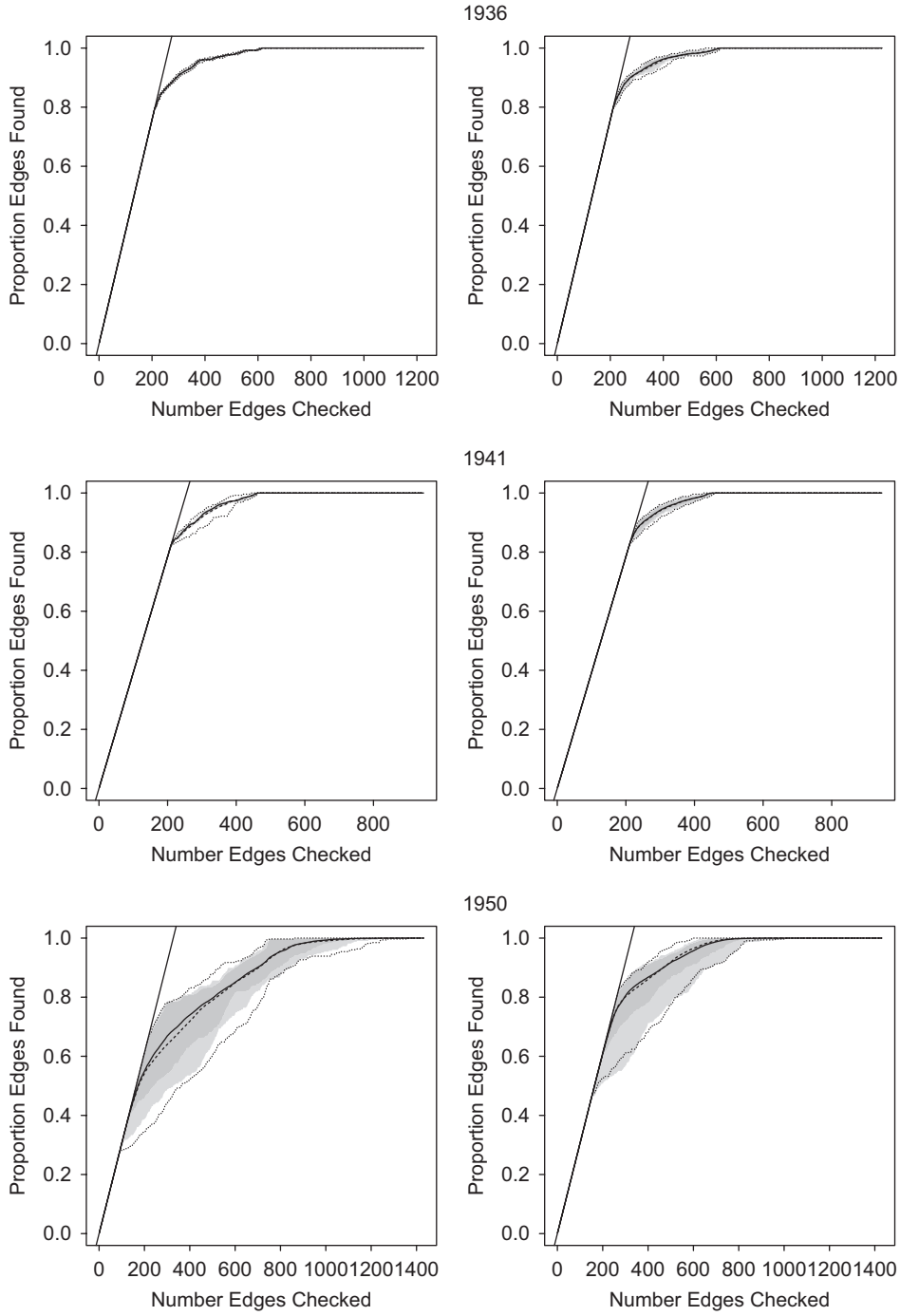

Fig. 5. Results for years 1936, 1941 and 1950.

in Fig. 9 shows the number of values still missing after using this simple imputation. Thus, there are some covariate values which are not observed for several years after the nation enters the graph.

Table 1 provides statistics on the years and shows the number of covariates that could be imputed using the temporal imputation. This analysis shows that while using past information can help fill in unknown covariates, there are countries which start out with unknown covariates, and so this alone cannot solve the problem. Also, there are trends

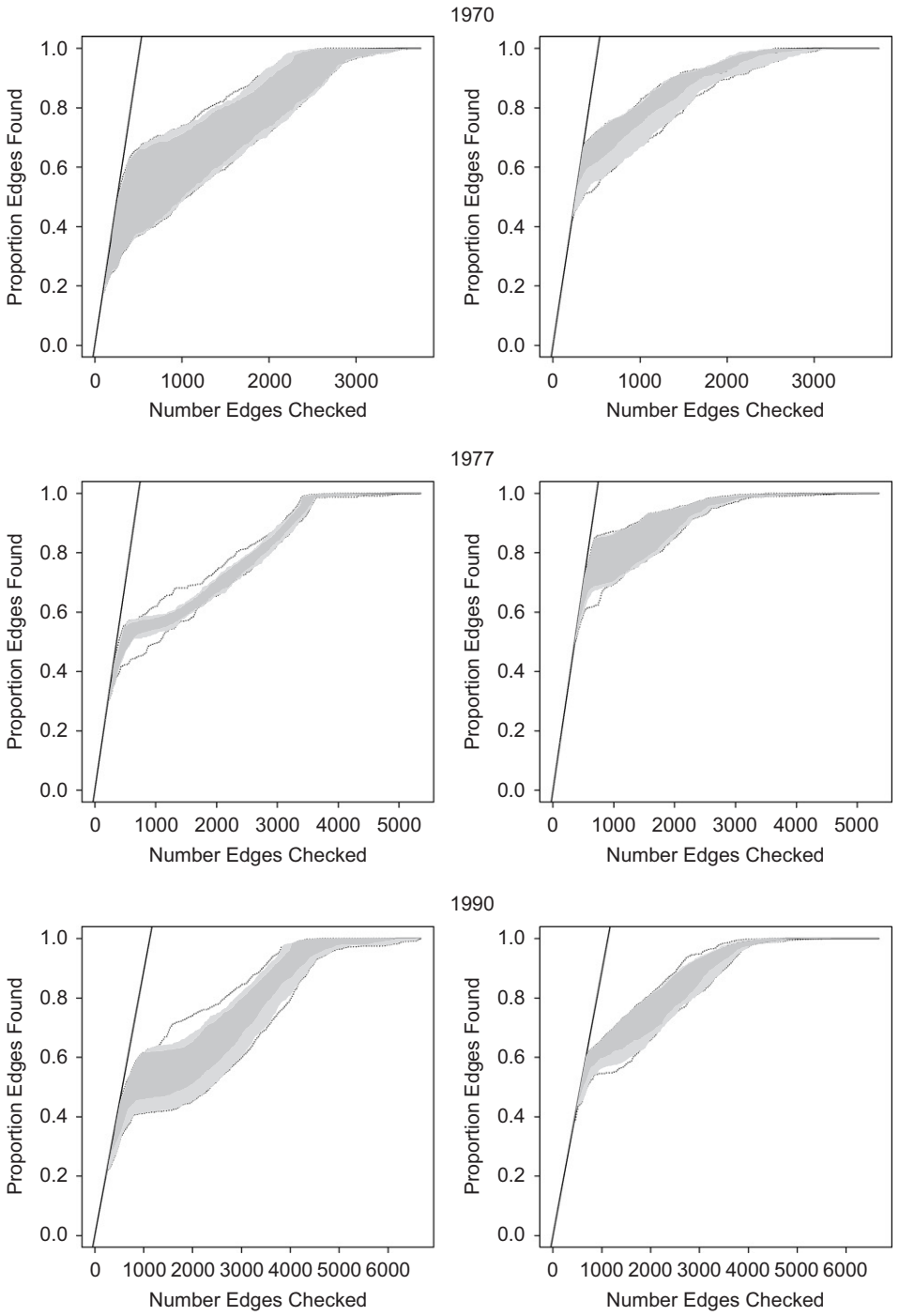

Fig. 6. Results for years 1970, 1977, and 1990.

in the covariate values, so even for those with past information, one should fit a model to the past data in order to predict the current values. This requires a model selection step, followed by a fitting of the model and imputation of the covariate value(s).

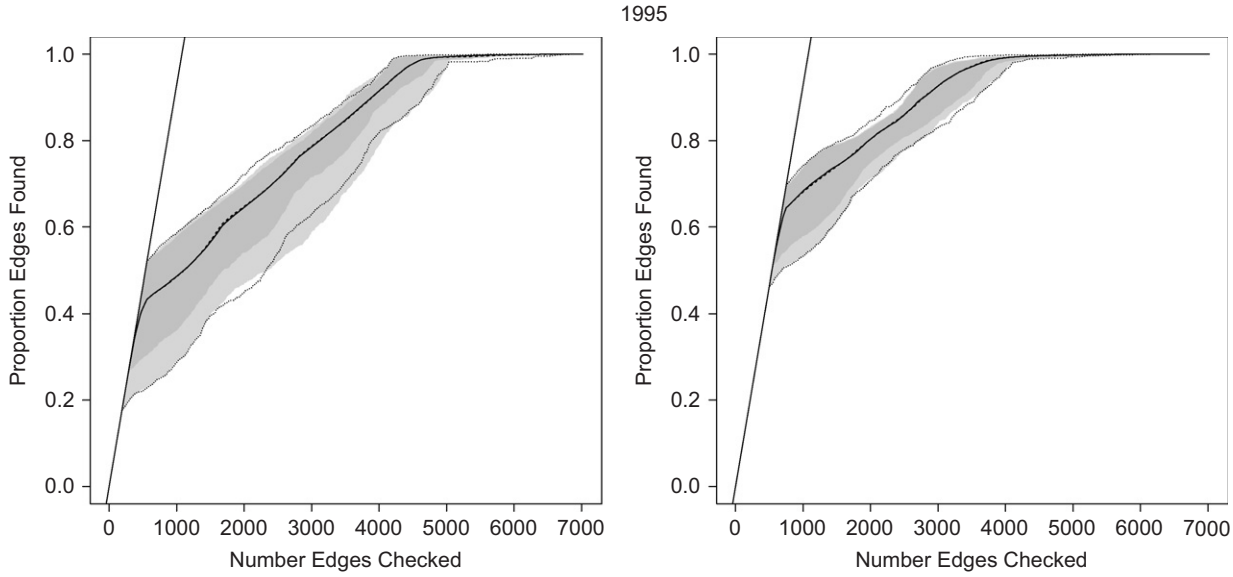

Fig. 7. Results for the year 1995.

Fig. 10 shows the first covariate, associated with military spending, for the United Kingdom. This shows long-term and short-term trends. Some short-term trends (around the two world wars, for example) are quite dramatic.

Instead of this time series approach, with the additional complication of the model selection on the time series of covariates, we ignored the temporal nature of the data and took a maximum likelihood approach to the missing data problem: we impute the missing values within a single graph as those for which the resulting likelihood is maximized. We hypothesize that by properly utilizing past information we could improve this imputation, but will not pursue this here.

To evaluate the use of covariate imputation, we consider those years in which the largest percent of values are missing: 1866 and 1995. The year 1941 also has a large number of missing attributes, but the prediction quality is so good for this year that there is little room for improvement. This graph has two components, one a clique of 21 nations (the US and its allies) and a loosely connected component corresponding to the rest of the world (see Fig. 1). The large clique explains the good performance of the edge prediction: edges missing at random in the clique are easy to predict.

The results for the two years in which attribute imputation was performed is shown in Fig. 11. The improvement provided by imputation is dramatic in 1995, where the largest number of missing covariates occurs.

4. Discussion

We have demonstrated how a simple model of networks, the dot product model, can be extended to utilize covariates, and can be used to predict the existence of unobserved edges, the idea being to prioritize resources expended to verify the unobserved edges. This paper dealt with edges missing at random (MAR); this may not be the best model for covert networks, where the nodes are actively attempting to hide their relationships. We have seen that adding appropriate covariates can improve the performance of the system, and have discussed some methods for dealing with missing covariates. Although we did not address this directly here, it is obvious that we could turn the probabilities around and rank the potential links that are most likely to truly be missing, although it may be harder to verify that a link is *not* present than it is to verify that it is.

The procedure has very different performance on the different years, and this is primarily due to the structure of the graphs. For example, in a graph consisting of cliques or near-cliques (such as 1941 in Fig. 1, see Fig. 5) prediction is relatively easy, since most potential edges in the dense regions are in fact true edges. In sparser graphs, covariates must be incorporated to correctly predict which of the missing edges is in fact a true edge.

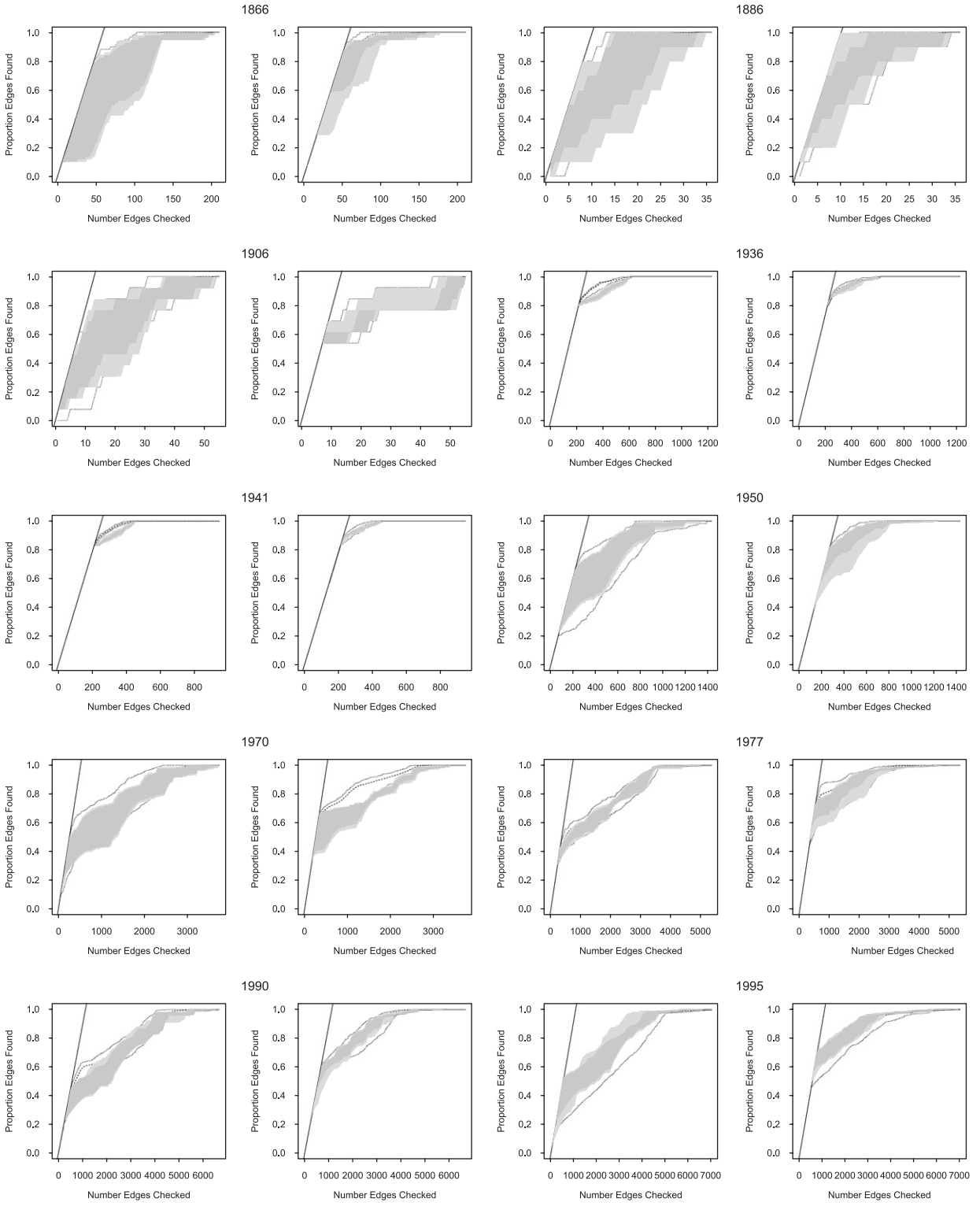

Fig. 8. Results for 100 2-fold crossvalidation runs for the 10 years studied. These are paired by year, with the first plot corresponding to the $K = 2$ model, and the second plot corresponding to the $K = 3$ model. In each case, the results corresponding to the incorporation of the three covariates is overlayed as dashed curves indicating the means and dotted curves indicating the envelope.

Table 1
Statistics for the years considered in the experiment

Year	$\|V\|$	$\|E\|$	#Missing covariates	% Missing	#Imputed
1866	11	10	8	24.2	0
1886	10	14	0	0	0
1906	11	13	0	0	0
1936	50	264	0	0	0
1941	44	255	26	19.7	25
1950	54	327	0	0	0
1970	87	517	3	1.1	2
1977	104	716	9	2.9	4
1990	116	1124	22	6.3	8
1995	119	1082	122	34.2	95

The number of non-isolated vertices and edges for the graphs are presented, as are the total number of missing covariates for the actors that year. The final column refers to the number of covariates that could be imputed via the temporal imputation method: fill in the most recent value, if any, for that covariate for that actor.

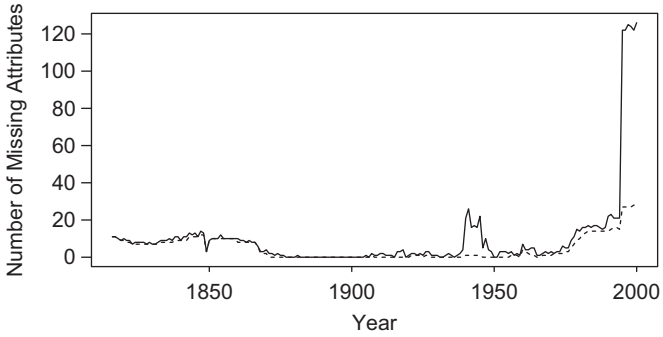

Fig. 9. Number of missing values in the covariates. The dashed curve shows the number of covariates missing after the simple imputation.

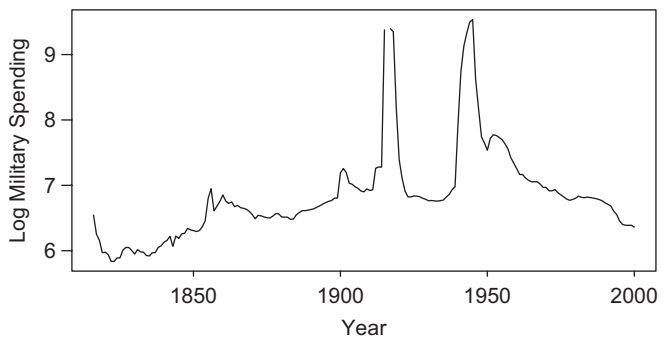

Fig. 10. Covariate values (log(military spending + 1) + 1) for the United Kingdom.

In a "real-world" system, one would recompute the probabilities after the edges had been verified. This would result in an iterative procedure, in which the algorithm proposes a list of potential relationships, a subset of these are checked directly (to the extent possible), and given this new information the algorithm is re-run to provide a new list of potential links. This would be easy to implement within the given framework.

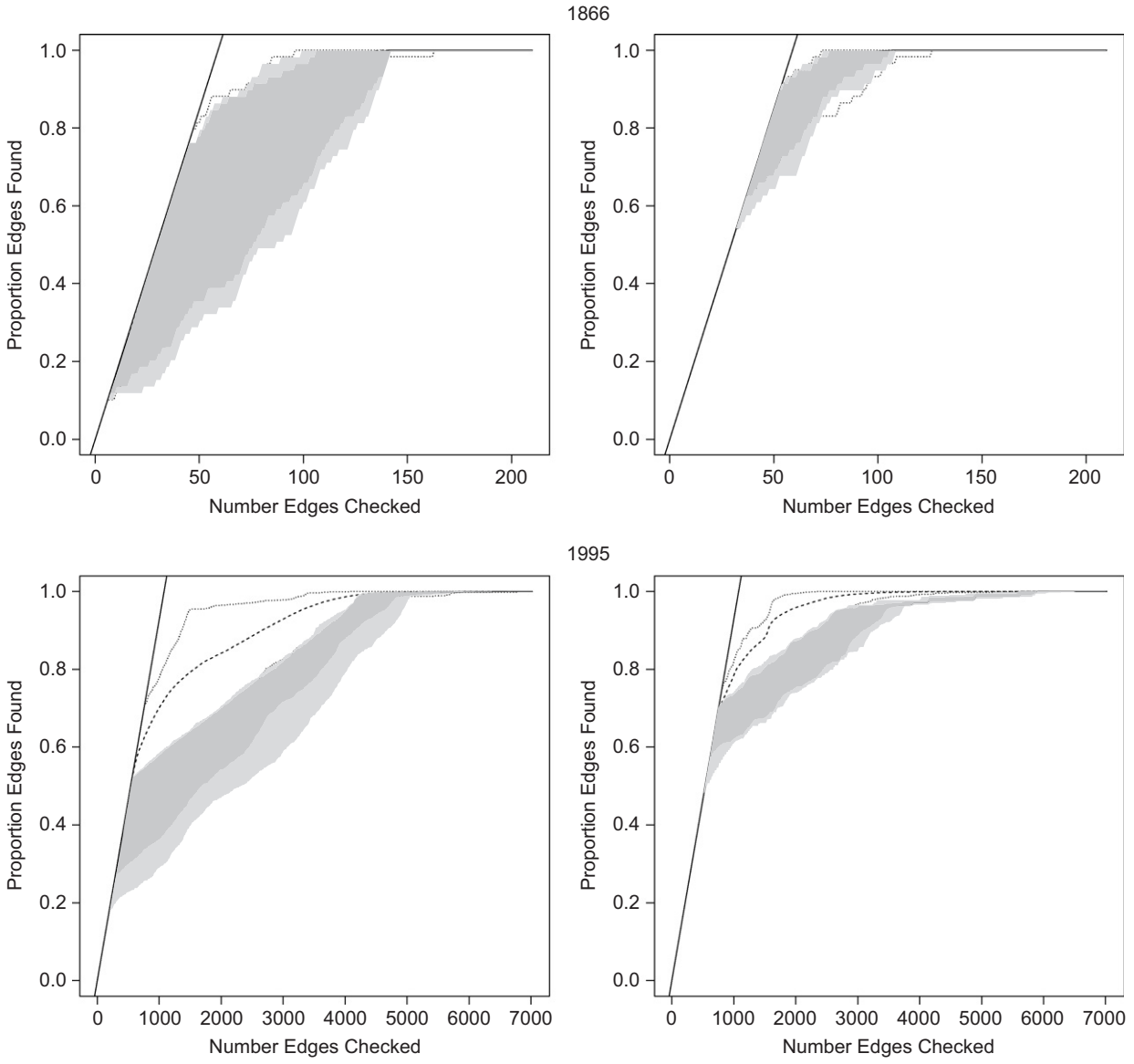

Fig. 11. Results for 100 4-fold crossvalidation runs for the 2 years studied using maximum likelihood for imputing missing covariates. The figures are as in Fig. 4. The years considered are 1866 and 1995.

We have not utilized the temporal nature of the data, except to note that simple temporal imputation of covariates may not be optimal (see Figs. 9 and 10 and Table 1). Time can be used to help predict edges, as well as covariates, and averaging edges across time can improve the estimation accuracy of the CRDPG model. Utilizing time series of graphs may also provide one method for approaching the problem of detecting missing nodes.

This procedure can be framed as an optimization of the reconstruction risk $P(g(v, w)|vw \notin E)$ where $g(v, w)$ is the edge prediction. The time series nature of the particular data set we investigated makes it a natural for a Bayesian approach. On can use the previous year's data in the formulation of a prior for the present year's graph, and choose the edges to maximize the posterior. This is an area of future research.

Our criterion for positing a link is based purely on the probability of the link. In many situations, one may have a statistical inference task, and may be interested in checking those links which would have the greatest impact on the statistical inference to be made. This is an area for future research.

References

Bollobás, B., 2001. Random Graphs. second ed. Cambridge University Press, Cambridge.
Gibler, D.M., Sarkees, M., 2004. Measuring alliances: the correlates of war formal interstate alliance data set, 1816–2000. J. Peace Res. 41, 211–222.
Hoff, P.D., 2005. Bilinear mixed-effects models for dyadic data. J. Amer. Statist. Assoc. 100 (469), 286–295.
Hoff, P.D., Ward, M.D., 2007. VMASC Statistics and Social Network Analysis Project Report.
Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. J. Amer. Statist. Assoc. 97, 1090–1098.
Karonski, M., Singer, K., Scheinerman, E., 1999. Random intersection graphs: the subgraph problem. Combin. Probab. Comput. 8, 131–159.
Kraetzl, M., Nickel, C., Scheinerman, E., 2007. Random dot product graphs: a model for social networks. Submitted for publication.
Marchette, D.J., Priebe, C.E., 2007. Modeling interstate alliances with constrained random dot product graphs. Comput. Statist., to appear.
Scheinerman, E., 2005. Random dot product graphs. Talk given at IPAM, ⟨www.ipam.ucla.edu/schedule.aspx?pc = gss2005⟩.
Scheinerman, E., Tucker, K., 2007. Modeling graphs using dot product representations. Comput. Statist., to appear.