Inference and Validation of Networks

Ilias N. Flaounas[1], Marco Turchi[2], Tijl De Bie[2], and Nello Cristianini[1,2]

[1] Department of Computer Science, Bristol University
Merchant Venturers Building, Woodland Road, Bristol, BS8-1UB, United Kingdom
[2] Department of Engineering Mathematics, Bristol University
Queen's Building, University Walk, Bristol, BS8-1TR, United Kingdom
http://patterns.enm.bris.ac.uk

Abstract. We develop a statistical methodology to validate the result of network inference algorithms, based on principles of statistical testing and machine learning. The comparison of results with reference networks, by means of similarity measures and null models, allows us to measure the significance of results, as well as their predictive power. The use of Generalised Linear Models allows us to explain the results in terms of available ground truth which we expect to be partially relevant. We present these methods for the case of inferring a network of News Outlets based on their preference of stories to cover. We compare three simple network inference methods and show how our technique can be used to choose between them. All the methods presented here can be directly applied to other domains where network inference is used.

Keywords: Network inference, Network validation, News Outlets network.

1 Introduction

Network Inference is a ubiquitous problem, found in fields as diverse as genomics, epidemiology or social sciences. Elements of a set (e.g., genes, people or news outlets) are connected by links that represent relations between them (e.g., co-expression, social contact, similar reporting bias, etc). While we can often observe the state of the network nodes, the underlying topology of the network is hidden, and must be inferred based on a finite set of observations of node-states.

Several methods have been proposed to infer this underlying network structure, in different communities and under different conditions. Examples include gene regulatory networks [1], biochemical regulatory networks[2] and protein interaction networks[3]. We focus on the general problem of testing, or validating, the result of this inference. Since often ground truth is missing, validation against related but different networks, or against networks inferred in different ways, is the only option or the only viable alternative to costly experiments.

In this paper, we present and study general methods to assess the results of Network Inference algorithms, from a statistical and machine learning point of view, and we demonstrate them on a challenging test case: the inference and

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 344–358, 2009.

validation of a network of News Outlets, based on content similarity information. All the principles and methods are however general, and can be applied to different domains.

We argue that network inference algorithms need to satisfy two key properties. First, the inferred network needs to be stable, meaning that networks inferred on independent data must be similar to each other. Second, it must be related to any available independent ground truth known or assumed to affect the network topology. Both these properties can be verified by testing if the inferred network is similar to a reference network.

For the first property, the reference network would be a network inferred based on independent data. For example for the network of News Outlets, we show that our network inference algorithm produces networks that are significantly similar to each other, when operating on independent datasets. This stability strongly indicates that the algorithm is capturing a signal, not noise.

For the second property, the reference network would be a network constructed based on independent ground truth data. For example for the network of News Outlets we show how the inferred network is significantly related to other—directly observable—networks of news outlets, such as those based on geographic, linguistic and media-type similarity.

Hence, both properties are verified by assessing if the inferred network is related to a reference network. More specifically, we want to verify if the inferred network is related significantly stronger to the reference network than a random network would be related to it. This is formalized in statistics by means of the key notion of statistical significance of a pattern, as expressed by a p-value. In order for this to be defined, we need to make two choices: a test statistic that quantifies how related the inferred network is to the reference network, and a null model for the inferred network.

The test statistic can be defined by quantifying the similarity of the inferred network to the reference network considered, and we will discuss various options for this similarity measure. To define the null model we will make use of two established approaches for random network generation. The choices we explore in this paper are exemplary, and other choices may be more appropriate in other applications. We will discuss the implications of these design choices, by comparing three different network inference algorithms for the News Outlet network inference application.

Finally, as a separate validation from a machine learning perspective, we investigate if we can predict the inferred network topology based on independent information. In particular, we show how Generalised Linear Models can be used to 'explain' the inferred network in terms of any known ground truth networks as discussed above.

Our approaches can be readily transferred to social sciences and genomics, where the availability of ground truth is the key problem when validating network inference.

2 Network Validation

The analysis of patterns found in data can generally be validated in two different ways: either by assessing their significance, or by measuring their predictive power. In the first case, we are interested in measuring the probability that a similar pattern could be found in randomly generated data. In the second case, we are interested in measuring the extent to which patterns found in a subset of the data, can be found in an independent subset of the data. Of course the two approaches have many relations, but in this study we will simply address them separately. We will call them respectively 'Hypothesis Testing' and 'Predictive Power' approach.

Notations. A network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ is comprised of the set of nodes $\mathcal{N} = \{N_1, N_2, ...N_n\}$ and the set of edges $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$, $n = |\mathcal{N}|$ is the total number of nodes of the network and $e = |\mathcal{E}|$ is the total number of edges.

2.1 Hypothesis Testing

The key idea of hypothesis testing is to quantify the probability that a the value of a test statistic evaluated on observed data could have been found also in random data. In our strategy to evaluate network inference algorithms, the test statistic is the similarity to a chosen reference network, and we denote it as $t_{\mathcal{G}_R}$ where $\mathcal{G}_R$ stands for the reference network. We denote as $\mathcal{G}_I$ the network inferred by the inference algorithm. The null hypothesis H_0 is that the inferred network is sampled from some underlying distribution. Hence, the hypothesis test boils down to quantifying the probability that a random network is at least as similar to a chosen reference network as the inferred network:

$$p = P_{\mathcal{G} \sim H_0}(t_{\mathcal{G}_R}(\mathcal{G}) \geq t_{\mathcal{G}_R}(\mathcal{G}_I)) \tag{1}$$

For most null hypotheses, it would be impractical to compute the p-value exactly. However, it can be reliably estimated by sampling a large number K of networks from the null hypothesis. Then the p-value is measured as the fraction of those for which the test statistic defined as the similarity to a reference network is smaller than that for the inferred network, or more precisely:

$$p \approx \frac{\#\{\mathcal{G} : t_{\mathcal{G}_R}(\mathcal{G}) \geq t_{\mathcal{G}_R}(\mathcal{G}_I)\} + 1}{K + 1} \tag{2}$$

If the p-value is small, this means that the inferred network is more similar to the reference network than expected by chance. Then the null hypothesis is rejected, as it is unlikely given the evidence. When the null hypothesis is rejected in this way, the alternative must hold. This means that the inferred network is significantly different from random networks sampled from the null hypothesis in its similarity to the reference network, supporting the inference method used to infer the network. In case that distances are used as test statistics instead of similarities, the inequality signs in p-value equations should be inverted. Often

a significance threshold α is chosen, and the null hypothesis is rejected if $p < \alpha$, where α is selected depending on the application.

In the following two subsections we will address the issue of selection of a test statistic and null models.

Test Statistics. As a test statistic we will use the similarity of the inferred network to a reference network, which we expect to be in some sense related to it. In this section we will discuss both the similarity measure and the choice of reference networks.

In literature several approaches have been proposed for the measurement of similarity between networks [4,5,6]. In the case of comparing two networks $\mathcal{G}_A$ and $\mathcal{G}_B$ that have the same set of nodes $\mathcal{N}$, one can compare the topological properties or their link structure. Perhaps the simplest comparison involves counting how many pairs of nodes have the same linkage status (connected or disconnected). The edges of each network can be considered as independent sets of elements and the comparison of networks is reduced to a comparison of sets of edges. Jaccard distance can be used to compare two sets as a measure of dissimilarity between them[7]. It is obtained by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

$$JD(\mathcal{E}_A, \mathcal{E}_B) = \frac{|\mathcal{E}_A \cup \mathcal{E}_B| - |\mathcal{E}_A \cap \mathcal{E}_B|}{|\mathcal{E}_A \cup \mathcal{E}_B|} \tag{3}$$

This quantity ranges between zero and one, with a value of zero indicating identical networks, and a value of one indicating no shared edges.

Of course one could define other measures, that consider less local properties, for example one could count how many triplets of nodes have the same connectivity status (in this way counting common network motifs [8]), or one could ignore the specifics of network topology, and focus on the distances between nodes represented by it. So a comparison of the all-pairs distance-matrix for each network could lead to a useful similarity measure.

In this study, as test statistic we will use the Jaccard distance from a reference network.

The choice of reference network is a very important one, as often in network inference applications we only have access to indirect evidence of the network topology (this being one of the key motivations for modern network inference). We have however often access to other networks (or sub-networks) for which the ground truth can be assumed to be known, and which we expect to be somewhat related to the network we are investigating.

If the data can be divided into independent sets, for example, we should assume that networks inferred on different parts of the data should be significantly similar. This can lead both to a bootstrap process, but also to the analysis of temporal data, as we will discuss in Sect. 3.3. Observing significant similarity to independently generated networks can provide strong support for a hypothesis, also in the case of networks generated with completely different types of data, as we will demonstrate using geographic, linguistic and media-type similarity, to test the significance of a network of news-media outlets.

Null Models. Every statistical test aims at answering the following question: what is the probability that a pattern like the one currently analysed is the result of chance? Of course this quantity (p-value) can be computed only after a random process has been specified, to formalise the notion of 'chance'. The Null Model has the crucial role of providing a baseline comparison to assess the significance of inferred patterns.

In the case of validating network patterns, we will need to specify a model of random network generation. If the observed similarity to a reference network is found also in randomly generated graphs, then we cannot conclude that we have found a significant pattern in the given data. In this study we will present two methods of random network generation, although many others are possible.

Erdös-Rényi Model
The first model is the celebrated $G(n, p)$ Erdös - Rényi model[9]. A graph of n nodes is generated by connecting nodes randomly. Two nodes have an independent probability p to be connected. This probability defines the density of the graph. Indeed the expected number of edges e is:

$$e = \binom{n}{2} p \tag{4}$$

and the distribution of the degree of any particular node N is binomial:

$$P(deg(N) = l) = \binom{n-1}{l} p^k (1-p)^{n-1-l} \tag{5}$$

Switching Randomisation
Although the Erdös-Rényi model is very natural and simple to analyse, it leads to topologies that are often very different from topologies observed in real world situations. For example, it does not exhibit the power-law in degree distributions that is often found in social networks. To remedy this, one can define a random-network generation model that—by construction—has the same degree distribution as the inferred network, and yet is randomly sampled from the space of possible networks. Such models can be created by a switching approach[10]. This method starts from a given graph and randomises it by switching edges between nodes. If the pairs of nodes A-B and C-D are connected, the model will switch the connections to create the edges A-D and B-C. The number of iterations is arbitrary but an adequate number is considered 100 times the number of edges [11].

2.2 Predictive Power

We are interested in the possibility of predicting the network topology based on other observable properties of the network. If some ground truth information about the inferred network is known, it is expected to be able to 'explain' the existence of some edges of the network. If more than one ground truth components are available this knowledge can be combined in order to improve the understanding of the inferred network. The combination of ground truth elements can be made using Generalized Linear Models (GLMs).

Generalized Linear Models. J. Nelder and R. Wedderburn introduced GLMs as a way to provide a unified framework for various non-linear or non-normal linear variations of regression[12]. GLM splits the model for the observed data Y_i into a random and a systematic component through a function called the link function. Under GLM Y_i is assumed to be generated from a distribution function of the exponential family [13]. The mean μ of the distribution depends on the independent variables, $\mathbf{X}$, through:

$$E(\mathbf{Y}) = \mu = g^{-1}(\eta) = g^{-1}(\mathbf{X}\beta) \tag{6}$$

where $E(\mathbf{Y})$ is the expected value of $\mathbf{Y}$; η is the linear predictor which is a linear combination of unknown parameters β; g is the link function; and the elements of $\mathbf{X}$ are typically measured by experimenters. The variance of the distribution is a function of the mean that can also follow the same exponential family distribution. The unknown parameters are easily estimated by maximum likelihood or other techniques.

Network Topology Prediction. The quality of the GLM models and the accepted ground truth components can be measured based on their power to predict the topology of the inferred network. Our aim is to measure the ability of the GLM model to predict the existence of an edge of the network. Using a methodology similar to this found in supervised classification we separate the network into a training and test sub-network. The training network is used to calculate the GLMs parameters. These parameters are combined with the accepted ground truth and are used to predict the structure of the test network. A generally accepted accuracy measurement is the Area Under Curve (AUC) based on the ROC analysis of the predictions on the test set. The separation into train and test sub-networks is performed multiple times under a cross-validation scheme in order to reduce bias.

3 Experimental Study

We will illustrate the validation methodology on the specific task of inferring the network of news outlets that are connected by the same bias in choosing stories to cover. This case study has many points in common with standard network inference tasks, for example gene regulation networks (while being easier to interpret): ground truth is not directly observed, side information is available, data is noisy, and so on. In this section we also introduce three increasingly complex network inference algorithms and use our methodology to compare their output.

3.1 Content-Based Inference of Media Outlets Network

In this application, we are interested in linking news outlets that have similar interests in choosing stories to cover. In this research a news story is defined as a set of news articles that cover the same event, practically found as a cluster.

We analyse a set of 1,017,348 articles gathered over a period of 12 consecutive weeks starting from October 1st, 2008, from 543 online news outlets, distributed over 32 different countries, in 22 different languages, including 7 different media types (e.g., newspapers, blogs, etc). This dataset was created as part of a separate project, which will not be discussed here[14]. While many of the outlets of interest offer their content in English language, we machine-translated the content of the others into English, by using Moses software [15,16].

Articles are preprocessed using stop-word removal, stemming and are vectorised using the TF-IDF representation[17]. They are then clustered in order to form the stories that will be used as the base for the network inference. The distance of two articles is measured using the cosine similarity[17]. The clustering algorithm identified on average 974 stories per day.

We used the Best Reciprocal Hit (BRH) clustering method, borrowed from the field of bioinformatics [18]. The choice of clustering algorithm is not central to the discussion of this study.

3.2 Three Network Inference Algorithms

We compared three network inference algorithms, all connecting pairs of nodes that have a sufficiently high level of similarity. While other inference methods are possible, we focused on this approach here for simplicity. Since we will use real valued similarity measures, we will also to choose a threshold in order to derive the linkage structure, and this threshold will control the density of the resulting graph.

We will assume we have an Outlet-by-Story matrix, indicating which outlets carried each given news story. There are 543 outlets, and 81,816 stories in total. Every outlet is hence described by an indicator vector in 'story-space'.

Method A. The simplest approach is to connect two outlets if they share some minimum number of stories. If the threshold is set to one, every pair of outlets that share at least one story are connected. In other words, the similarity measure between outlets is the scalar product between their indicator-vectors in story-space. This approach can easily lead to very dense networks since many stories are shared by the majority of outlets.

Method B. A more sophisticated approach would apply weights to the candidate edges of the network. A popular weighing scheme is based on the TF-IDF. Under this scheme each outlet correspond to a document and each story to a term. The frequency of story j that belong to outlet k is

$$f_j^k = \frac{s_j^k}{s^k} \tag{7}$$

where the nominator is the number of times the story appears to the outlet k and s^k is the total number of stories of outlet k. The corresponding inverse outlet frequency i_j^k is defined as

$$i_j^k = \log \frac{n}{n_j} \tag{8}$$

where n is the total number of outlets and n_j is the number of outlets that have story j. Thus, a vector of size J, that is the total number of different stories, is assigned to each outlet, one weight for each story:

$$w_j^k = f_j^k \cdot i_j^k, j = 1, 2..J \tag{9}$$

The similarity of two outlets N_a and N_b can now be measured as their cosine similarity:

$$sim(N_a, N_b) = \sum_{t=1}^{J} w_t^a w_t^b \tag{10}$$

Method C. Another method which is similar to the previous one is weighting each story with a weight f_j based on the frequency of the story, independently of the outlet that publish it:

$$f_j = \frac{1}{n_j} \tag{11}$$

where n_j is the number of outlets that have story j. Stories that are found in the majority of media receive a small weight and stories found in few media receive higher weight. The maximum weight is $1/2$ since we consider as stories clusters that have articles of at least two different outlets, and the minimum weight is $1/n$. If we normalise the above measure to the range of zero to one we get

$$f'_j = \frac{2(n - n_j)}{(n-2) \cdot n_j} \tag{12}$$

where n is the total number of outlets and we consider two as the minimum number of outlets that can belong to a cluster. This way measure of similarity between two outlets N_a and N_b is defined as:

$$sim'(N_a, N_b) = \frac{\sum_{t=1}^{J} f'_t y_a(j) y_b(j)}{\sum_{t=1}^{J} y_a(t) y_b(t)} \tag{13}$$

where $y_k(j)$ is one if outlet k has story j and zero otherwise.

3.3 Results

In this section we present the application of our methodology for inferring and validating the News Outlets network. We show that the network presents stability in time using independent datasets, that using some ground truth knowledge we can select the appropriate inference algorithm, and that finally we can predict the network structure.

Stability in Time. Figure 1 presents the stability of the network for the 12 consecutive weeks. The dataset of each week is independent of the data of the other weeks and the first week is used only as reference network. The threshold

of the three network inference methods was set to produce networks of the same density of ~ 5000 edges per week.

To determine the stability of the network inference algorithm, we test whether an inferred network's similarity to the inferred network from the previous week is significant. In order to do this, we carry out two hypothesis tests: one for each of the possible null models discussed in 2.1. As test statistic we used the Jaccard Distance. We sampled $K = 1000$ networks and estimating the p-value as the fraction of those for which the Jaccard distance with the reference network from the previous week is smaller than for the inferred network.

The networks inferred by each method are significantly similar ($p < 0.001$) to those inferred the previous week with the same method, under both null models.

Comparison to Related Networks. In this test, we compare the network inferred by each of the three Methods, with three other networks, obtained using independent (but possibly related) information. The first one (Location Network) linking outlets with the same geographic location; the second one (Language Network) linking outlets written in the same language; the third one (Media Type) liking outlets of the same type (e.g., newspaper, magazine, broadcast, blog, etc). These networks of outlets are formed of several disjoint cliques, and we expect some of them to relate to the news-choice preference of an outlet. Clearly, a story that is important and publishable for UK media may be uninteresting to the French media. Language is also an important factor, independently of the location of the outlet. For example, we measured that the number of stories that mention the word 'Pope' in Spanish-language media in the USA is three times larger than in English-language media in the same country. About the influence of media type, it is worth mentioning that certain stories may reported in blogs before they appear in mainstream traditional media.

Figure 1 presents the comparison of the content-based network inferred by Methods A, B, C to the 'Location' network. In this case only the Methods B and C are significant with $p < 0.001$. Figure 2 presents the case of the 'Language' where only Method C yields significant patterns ($p < 0.001$) over all weeks' datasets. Finally Fig. 2 illustrates the 'Media-Type' case where Method A and Method C yield significant results ($p < 0.001$). Only method C present significant results for all reference networks over all independent datasets and the two null models that were used to make comparisons. Note that although the distances between networks seem relative small, they are highly significant.

Selecting Inference Method. The selection of the inference method will be made based on their ability to create significant results. We selected a significance level of 0.001 and based our decision on this. Table 1 compares the three methods for the 11 weeks' independent datasets (the first week was used only as reference network). The numbers represent the number of weeks that the methods presented significant results with $p < 0.001$ for the different reference networks and the two null models. Only Method C presents significant results for all the performed tests and datasets, performing at least as good as the two

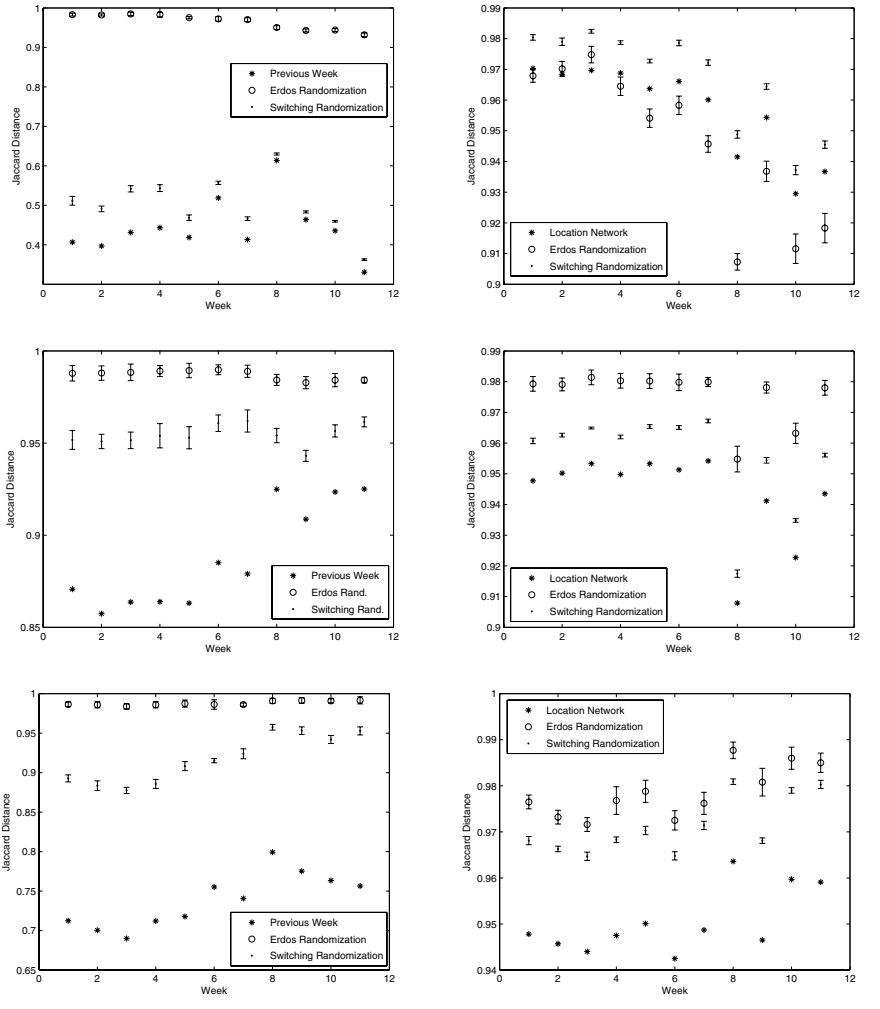

Fig. 1. Network stability on sequential weeks (on the left) and 'Location' reference network (on the right) for the three network reconstruction methods. Errorbars are $+/-$ 3 standard deviations over the mean value.

other methods investigated on all tests carried out. We can therefore conclude that Method C is better than both Methods A and B.

In Figure 3 we report the network of the media outlets obtained with Method C. To the best of our knowledge, this is the first map of this kind to be published. The visualisation of the network was made by using the Cytoscape software [19].

Prediction of Edges. We investigated the ability of prediction of an edge of media outlet network based on the GLM analysis and the three available ground truth reference networks. For the GLM analysis we adopted the normal distribution for Y_i and the identity link function where $\mu = X\beta$. The accuracy

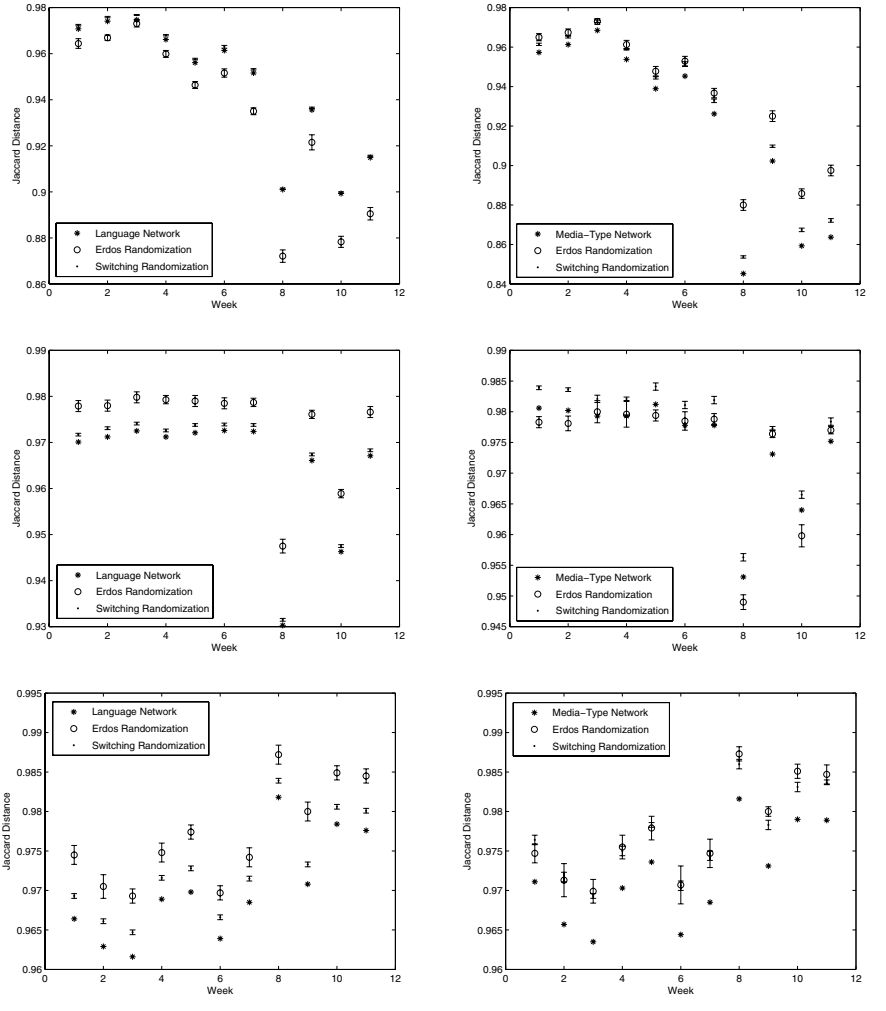

Fig. 2. Comparison to the 'Language' (on the left) and 'Media-Type' (on the right) reference networks. Errorbars are +/− 3 standard deviations over the mean value.

Table 1. The number of weeks that each Method presented significant results with $p < 0.001$. ER stands for the Erdös and SR for the Switching Random Graphs.

	Previous week		Location		Language		Media-Type	
	ER	SR	ER	SR	ER	SR	ER	SR
Method A	11	11	1	11	0	9	11	11
Method B	11	11	11	11	11	11	2	11
Method C	11	11	11	11	11	11	11	11

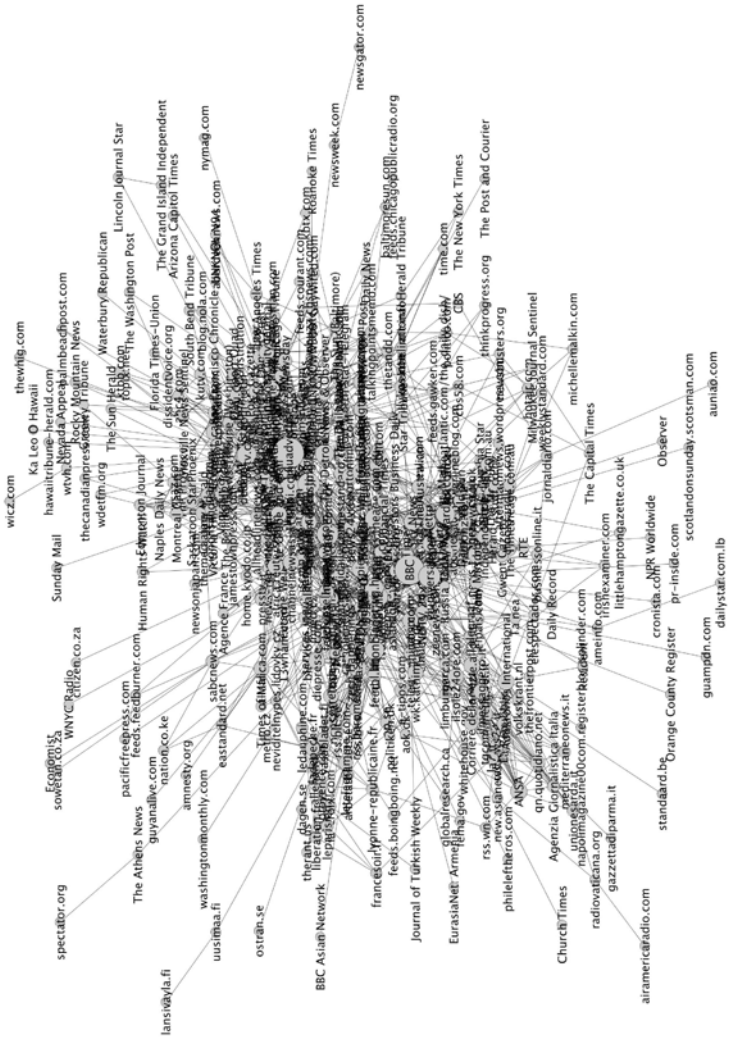

Fig. 3. A snapshot of the News Media Network. A high threshold is set to produce more sparse graph for visualisation reasons. This graph contains 351 nodes and 1612 edges. Singletons are omitted. The node sizes are proportional to their degrees.

was measured as the AUC for a 100-fold cross-validation scheme for different densities of the inferred network. For each inference method two figures are presented: The first illustrates the accuracy of each ground component if it was used by itself for predictions compared to using all three, and the second the accuracy of using pairs of components compared to using all three. Figure 4 present the edge prediction results for the three network inference methods and under different scenarios: Using all three ground truth networks combined, using each one of them separately and using all pairwise combinations of them. The

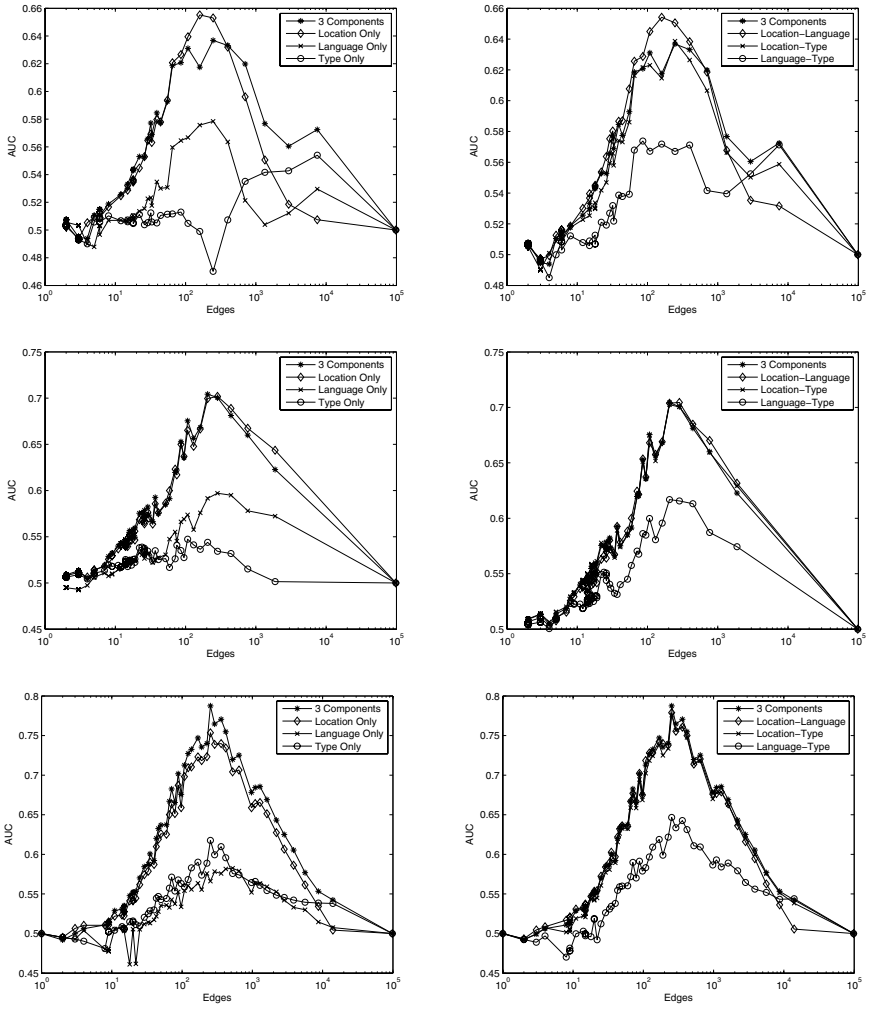

Fig. 4. AUC accuracy for edge prediction using Method A on the top row, Method B in the middle row and Method C in the bottom row, based on GLM analysis over different network densities

best prediction accuracy, 77.11%, over all different network densities, is reached using all three ground truth networks and the Method C.

4 Conclusions

The validation of network inference results in terms of statistical assumptions or in terms of related networks, indicates how to handle the common case where ground truth is difficult to obtain. Concepts from statistical testing can directly

provide a framework for assessing and comparing results, algorithms and also datasets.

Importantly, when one can distinguish in a principled way between two algorithms, then one can also search the hypothesis space for the best possible network. Future work in this direction will include the design of network inference algorithms that directly optimise the stability and significance of the output, instead of just choosing between existing heuristic algorithms.

Acknowledgements. The authors want to thank Omar Ali and Phil Naylor, respectively for their contribution in data gathering and computer infrastructure support and the entire 'Pattern Analysis and Intelligent Systems' group at the University of Bristol for discussions. This work is partially supported by the European Commission through the PASCAL2 Network of Excellence (FP7-216866), and the IST project SMART (FP6-033917). Ilias Flaounas is supported by Alexander S. Onassis Public Benefit Foundation and Nello Cristianini is supported by a Royal Society Wolfson Merit Award.

References

1. D'haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics 16, 707–726 (2000)
2. Ma'ayan, A.: Insights into the Organization of Biochemical Regulatory Networks Using Graph Theory Analyses. J. Biol. Chem. 284, 5451–5455 (2009)
3. Paris, L., Bazzoni, G.: The Protein Interaction Network of the Epithelial Junctional Complex: A System-Level Analysis. Mol. Biol. Cell 19, 5409–5421 (2008)
4. Pelillo, M.: Replicator Equations, Maximal Cliques, and Graph Isomorphism. Neural Computation 11(8), 1933–1955 (1999)
5. Bunke, H.: Error Correcting Graph Matching: On the Influence of the Underlying Cost Function. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 917–922 (1999)
6. Fernández, M.-L., Valiente, G.: A graph distance metric combining maximum common subgraph and minimum common supergraph. Pattern Recognition Letters 22, 753–758 (2001)
7. Jaccard, P.: Ètude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Societ. Vaudoise des Sciences Naturelles 37, 547–579 (1901)
8. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon1, U.: Network Motifs: Simple Building Blocks of Complex Networks. Science 298, 824–827 (2002)
9. Erdös, P., Rényi, A.: On Random Graphs. Publicationes Mathematicae 6, 290–297 (1959)
10. Rao, A.R., Jana, R., Bandyopadhya, S.: A Markov chain Monte Carlo method for generating random (0, 1)-matrices with given marginals. Indian J. of Statistics 58, 225–242 (1996)
11. Milo, R., Kashtan, N., Itzkovitz, S., Newman, M.E.J., Alon, U.: On the uniform generation of random graphs with prescribed degree sequences (2003) Arxiv condmat/0312028

12. Nelder, J., Wedderburn, R.: Generalized Linear Models. Journal of the Royal Statistical Society 135 Series A (General), 370–384 (1972)
13. McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall, London (1989)
14. Turchi, M., Flaounas, I., Ali, O., De Bie, T., Snowsill, T., Cristianini, N.: Found In Translation. In: Buntine, W., et al. (eds.) ECML/PKDD 2009. LNCS, vol. 5781. Springer, Heidelberg (2009),
http://patterns.enm.bris.ac.uk/publications/found-in-translation
(accepted for publication)
15. Koehn, P., Hoang, H., et al.: Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting-Association for Computational Linguistics 45 (2007)
16. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54. Association for Computational Linguistics, Morristown (2003)
17. Liu, B.: Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg (2007)
18. Hirsh, A., Fraser, H.: Protein dispensability and rate of evolution. Nature 411, 1046–1049 (2001)
19. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, I.T.: Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13, 2498–2504 (2003)