

Predicting Relations in News-Media Content among EU Countries

Ilias N. Flaounas, Nick Fyson , Nello Cristianini
Intelligent Systems Laboratory, University of Bristol, Bristol, BS8 1UB, UK
<http://intelligentsystems.bristol.ac.uk>
{[ilias.flaounas](mailto:ilias.flaounas@bristol.ac.uk), [nick.fyson](mailto:nick.fyson@bristol.ac.uk)}@bristol.ac.uk, nello@support-vector.net

Abstract—We investigate the complex relations existing within news content in the 27 countries of the European Union (EU). In particular we are interested in detecting and modelling any biases in the patterns of content that appear in news outlets of different countries.

We make use of a large scale infrastructure to gather, translate and analyse data from the most representative news outlets of each country in the EU. In order to model the relations found in this data, we extract from it different networks expressing relations between countries: one based on similarities in the choice of news stories, the other based on the amount of attention paid by one country to another. We develop methods to test the significance of the patterns we detect, and to explain them in terms of other networks we created based on trade, geographic proximity and Eurovision voting patterns. We show that media content networks are 1) stable over time, and hence well defined as patterns of the news media sphere; 2) significantly related to trade, geography and Eurovision voting patterns; 3) by combining all the relevant side information, it is possible to predict the structure of the media content network.

In order to achieve the above results, we develop various pattern analysis methods to quantify and test the non-metric, non-symmetric pairwise relations involved in this data. These methods are general and likely to be useful in many other domains.

I. INTRODUCTION

Many aspects of our society are shaped by the contents of our news-media system, yet a large scale systematic investigation of its content is difficult to perform, and as a result its global features are not very well understood. There are various orders of reasons for this: the screening of vast amounts of multilingual data requires automation of all steps of the process, from acquisition to translation to content analysis. The modelling of the relations extracted from the data requires a very diverse set of tools.

In this paper we examine relations existing between the news outlet (newspapers, magazines, broadcast media etc) content in all EU countries over a period of six months, and we compare them with a number of other relations between the same countries (geographic, cultural, commercial). By using network analysis we show how news content patterns can be explained in terms of these other relations.

The methods we present in this study are general, and can be applied to other problems where machine learning and pattern recognition need to be performed on data consisting of non-metric and non-symmetric relations between pairs of items.

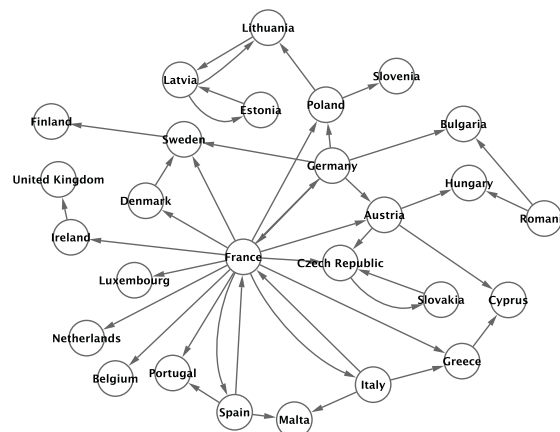


Fig. 1. ‘Citations’ network for July 2009. Country A points to country B if B is mentioned ‘frequently’ in news outlets of country A.

II. DATA GATHERING

We generated and studied two networks based on the content of news articles of European media outlets. Figure 1 illustrates a network of the EU countries where an edge from country A to country B is created if the number of citations from country A to country B is significantly greater than the average number of citations country A gives across all countries. Figure 2 presents an alternative network, where connection between countries is made based on the criterion that their respective media outlets cover the same stories [1].

The dataset we used includes six months from May 1st 2009 to October 31, 2009 in 255 EU outlets, selected in a way to represent the main outlets of each country (as captured by Alexa.com ranking scores). Of these outlets, we used only the news items advertised in the main RSS feed of their website. This makes a total of 1,419,251 news items, in 22 EU languages. This data was then translated to English based on Moses software [2]. The infrastructure to generate this dataset was created in a separate project [3].

A. News Content Networks

On the translated data we inferred two different types of networks, both of them directed by construction:

1) ‘Citations’ network: The first type is based on how often a country’s media mentions one of the other EU countries. We link country A to country B if the fraction of citations from

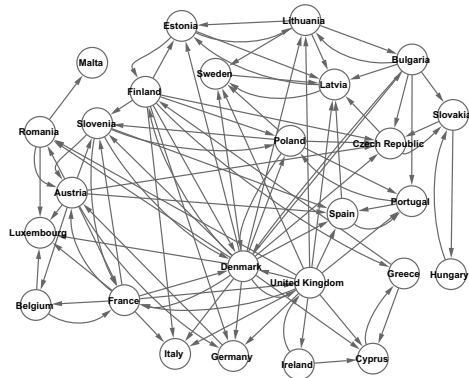


Fig. 2. 'Coverage similarity' network' for July 2009. Country A points to country B, if B has more than average interest in the same news stories with country A.

A to B is significantly larger than the average citation count from A across all countries. An example of this network as inferred for July 2009 is presented in Fig. 1, for a threshold of one standard deviation above mean value.

2) 'Coverage similarity' network: The second network is based on a clustering of the news-items into 'news stories', i.e. sets of articles covering the same event. The news items of each day were clustered based on bag of words similarity, giving each country a set of stories with which it was associated. Each entry in the adjacency matrix was then a measure of the overlap of story clusters for the countries in question. The same thresholding strategy is applied as before, resulting in a different threshold for each country. An example of this type of network is shown in Fig. 2, for July 2009 and a threshold of one standard deviation above mean.

B. Reference Networks of Country Relations

We used three reference networks of EU countries that are based on non-textual information and can be considered as ground truth. These are based on geographic, trade and cultural relations between countries:

1) 'Trade': Data is from United Nations Statistics Division - Commodity Trade Statistics Database¹. We take the total of all trade between the respective countries in the year 2008. Normalisation is performed by dividing each row by its total, such that edge weights are the fraction of the particular country's trade that is directed towards each other country. This network is illustrated in Fig 3.

2) 'Song Contest': Data is from Eurovision song contest for years from 1957 to 2003², with the number of points awarded by each country to each other country summed over the whole time period. Normalisation is performed by dividing each entry by the row total, such that the edge weights indicate

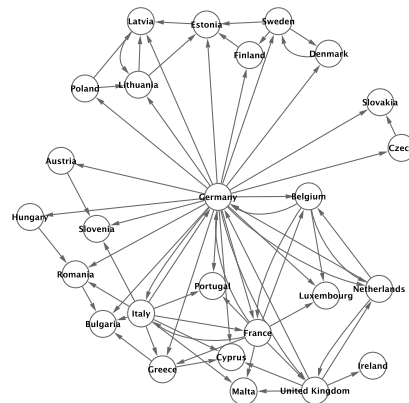


Fig. 3. Trade Network: a country points to another country if this represents a share of trade significantly larger than the average trading partner for that country.

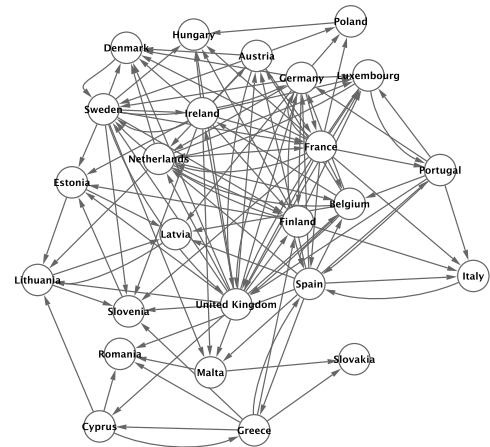


Fig. 4. Song Contest voting Network: a country points to another country if this represents a voting preference significantly larger than the average votes for that country.

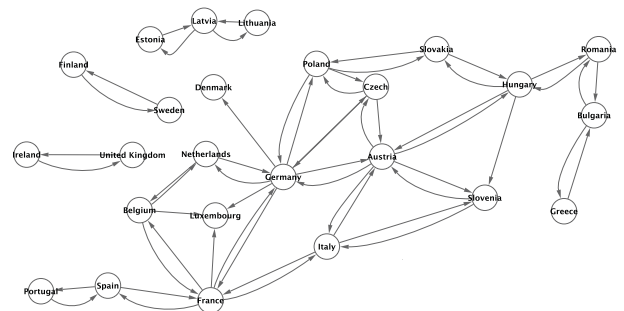


Fig. 5. Borders network: a country points to another country if this represents a share of borders significantly larger than the average borders length for that country. Singletons are omitted.

the fraction of total points awarded by the country in question to each other country. Countries present in the voting data but not in the current EU countries list were removed prior to normalisation. This network is illustrated in Fig 4.

3) 'Borders': An adjacency matrix was constructed based

¹COMTRADE: <http://comtrade.un.org/db>

²Eurovision Song Contest: <http://www.eurovision.tv>

on whether two countries share land borders. Normalisation was again performed by dividing by the row total, such that the edges indicated the fraction of a country's borders that are shared with each other country. This makes a directed network, as illustrated in Fig 5.

III. METHODS

We formed networks representing pairwise relations among countries in a very simple way, since the focus of this paper is on network analysis, rather than on network inference.

We address the following key questions: Q1) is there a well defined network of relations between countries reflecting the content of their news outlets? (e.g. similarity between countries, or attention levels from one country to another) Q2) can these networks be explained in terms of other relations? (e.g. existing commercial, geographic and cultural relations?)

Similar questions can easily be encountered in other domains. For example, in bioinformatics one may wonder if a metabolic network can be explained in terms of a gene regulation network, or if the biological networks in an organism are significantly similar to the networks in another organism [4]. We feel that developing general approaches to these questions is important.

The first question (Q1) is addressed by developing a statistical test for network similarity, consisting of a similarity measure and of a null model for network generation. The aim is to demonstrate significantly high similarity between content networks based on different time periods, and hence show that these networks capture a genuine property of the news media system.

The second question (Q2) is addressed by developing a simple machine learning algorithm that can predict the relations in content between countries based on relations of a different nature. If this prediction is possible then one can start explaining patterns in media content with other factors. For this we use Generalised Linear Models (GLM).

A. Statistical Testing (Q1)

In testing the significance of the patterns we observe, we calculate a p -value: the probability that a given pattern is due to chance, given stated assumptions about random generation of data. This involves defining a 'test statistic' and a 'null model', in this case a similarity measure between networks and a model for the stochastic generation of networks. We then determine the probability that the observed level of similarity was observed through chance alone, or whether it is indicative of a significant underlying pattern. If the p -value is small we may conclude that there is statistically significant similarity between the networks under comparison. Usually it is impractical to compute the p -value exactly. However, it can be estimated by sampling a large number K of networks from the null model and calculating the test statistic for each instance. The p -value is then simply the fraction of those for which the test statistic $t_{\mathcal{G}_R}$ (defined as the similarity to a reference network \mathcal{G}_R) is smaller than that for the inferred

network \mathcal{G}_I :

$$p \approx \frac{\#\{\mathcal{G} : t_{\mathcal{G}_R}(\mathcal{G}) \leq t_{\mathcal{G}_R}(\mathcal{G}_I)\} + 1}{K + 1} \quad (1)$$

1) *Test Statistic*: In literature several approaches have been proposed for the measurement of similarity between networks [5]–[7]. Since the networks under examination in this paper have the same set of nodes (corresponding to the 27 countries of the EU) we can make use of a very simple similarity measure as our test statistic. We use a directed version of 'Jaccard Distance'. Jaccard distance can be used to compare two sets of edges, as a measure of dissimilarity between them. It is obtained by taking the difference of the sizes of the union and the intersection of two sets, and dividing by the size of the union:

$$JD(\mathcal{E}_A, \mathcal{E}_B) = \frac{|\mathcal{E}_A \cup \mathcal{E}_B| - |\mathcal{E}_A \cap \mathcal{E}_B|}{|\mathcal{E}_A \cup \mathcal{E}_B|} \quad (2)$$

This quantity ranges between zero and one, with a value of zero indicating identical networks and a value of one indicating no shared edges. The directed version of this measure takes into account the directionality of the edges between the two networks under consideration. An edge between two nodes that has both directionalities counts as two separate edges. Only if both networks share an edge between the same nodes with the same directionality will it count towards the intersection component of the equation.

2) *Null Models*: We make use of two simple randomisation strategies to generate networks with similar properties to the network under examination. The first one is the $G(n, p)$ Erdős - Rényi model [8]. 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. The Erdős-Rényi model creates network topologies that are very different from the topologies observed in real-world situations, e.g. it does not preserve the power-law that is often found in the node degree distribution of social networks [9].

The 'switching' randomisation strategy described gives rise to networks with the same degree distribution but randomised topology [10]. This method starts from a given graph and randomises it by switching edges between nodes. For example, if edges $A \rightarrow B$ and $C \rightarrow D$ are present, the model will switch the connections to create the edges $A \rightarrow D$ and $B \rightarrow C$. The number of swaps is arbitrary but an adequate number can be considered 100 times the number of edges.

B. Generalised Linear Models (Q2)

We analyse the news content networks as combinations of the three reference networks using Generalised Linear Models. GLMs were introduced by J. Nelder and R. Wedderburn as a unified framework for various non-linear or non-normal linear variations of regression [11], [12]. GLM assume the observed data, that is network edges in our framework, Y_i can be splitted into a random and a systematic component through a function called the 'link function'. Data is assumed to be generated

from a distribution function of the exponential family. The mean μ of the distribution depends on the independent variables, \mathbf{X} , through:

$$E(\mathbf{Y}) = \mu = g^{-1}(\eta) = g^{-1}(\mathbf{X}\beta) \quad (3)$$

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 can be estimated by maximum likelihood or other techniques.

The quality of the GLMs and the accepted reference networks can be measured based on their power to predict the topology of the inferred networks. Our aim is to measure the ability of the GLM 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 [13], [14]. The separation into train and test sub-networks is performed multiple times under a cross-validation scheme.

IV. RESULTS

We show that the media content networks that we inferred, namely ‘Citations’ and ‘Coverage-similarity’, are stable over time and hence well defined as patterns of the news media sphere; next that they are significantly related to the ground truth networks of trade, geography and song contest voting patterns; and finally that by combining all this independent relevant side information, it is possible to predict their topology.

A. Stability in time

We have found that both ‘Citations’ and ‘Coverage-similarity’ networks are significantly stable compared to randomly generated networks ($p < 0.001$ for both null models – 1000 random networks were generated per null model). Figure 6 illustrates a comparison of distances between sequential ‘citations’ and ‘Coverage-similarity’ networks. For each case, each month a network is inferred and it is compared to three other networks: The network of the previous month, a random network generated using the Erdős model, and a random network as a result of switching randomization. The density of the Erdős random network was chosen to be similar to that of the inferred network.

B. Relation to ground truth

We examined the relation of both news content networks to the three reference networks. Figure 7 illustrates the distances of the ‘Citations’ network to ‘Trade’, ‘Song Contest’ and

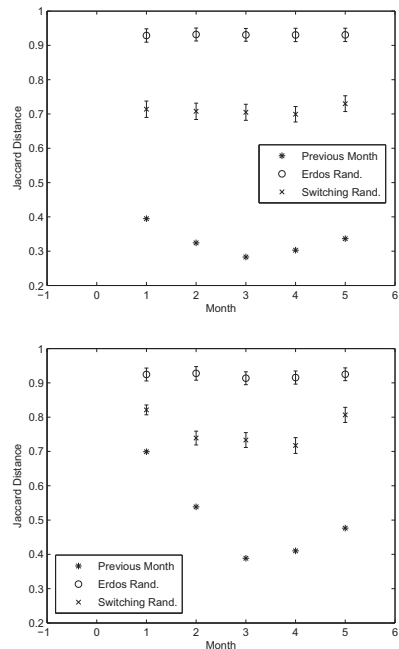


Fig. 6. Stability of the ‘Citations’ (top) and ‘Coverage-similarity’ (bottom) networks: comparison of sequential months.

‘Borders’ networks respectively. All comparisons are significant ($p < 0.001$) for both of the two null models. Figure 8 presents the corresponding distances for the ‘Coverage-similarity’ network.

C. Topology prediction

We used GLMs with normal distribution in order to predict the two networks topology. We divided the networks edges in train and test sets under a 10-fold cross validation scheme. Figure 9 presents the results for the ‘Citations’ networks for the six month period of study. Each month four different networks were inferred for different density thresholds (the mean value, one, two and three standard deviations above mean). In each plot we present the AUC accuracy of prediction using all three reference networks and each one of them separately for comparison. It can be seen that prediction accuracy reached up to 89.77% for August using the information of all three reference networks. Most of the times the best single predictor is the ‘Trade’ network, followed by the ‘Borders’ network and then by the ‘Song contest’ network.

Figure 10 illustrates the topology prediction results for the ‘Coverage-similarity’ network. We used a 10-fold cross validation scheme and three different network densities (the mean, one and two standard deviations above mean). The best results of 77.74% are achieved using all three reference networks for June.

V. CONCLUSIONS

Complex non-metric relations are commonplace in domains as diverse as sociology, molecular biology and epidemiology,

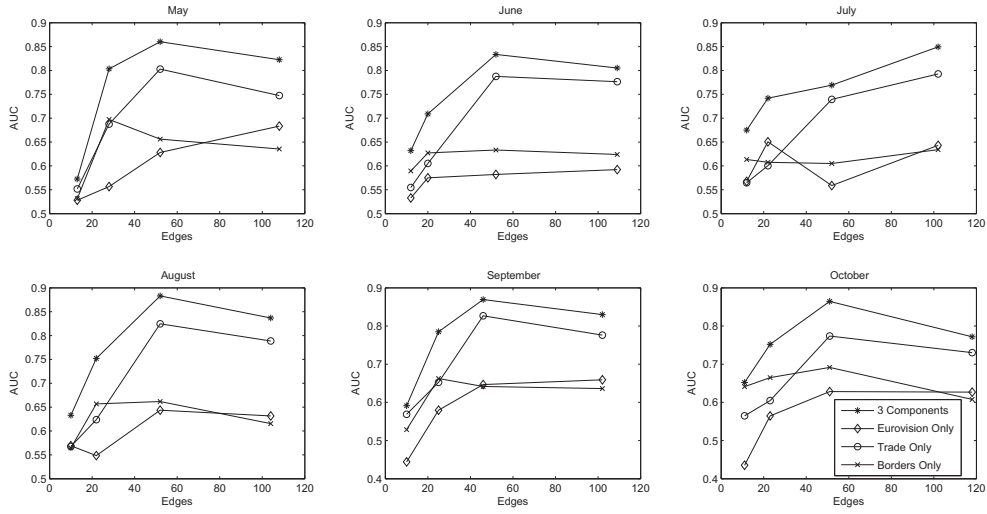


Fig. 9. Prediction AUC accuracies, under a 10-fold cross-validation scheme, for the ‘Citations’ network for four different network densities over the six months period of study.

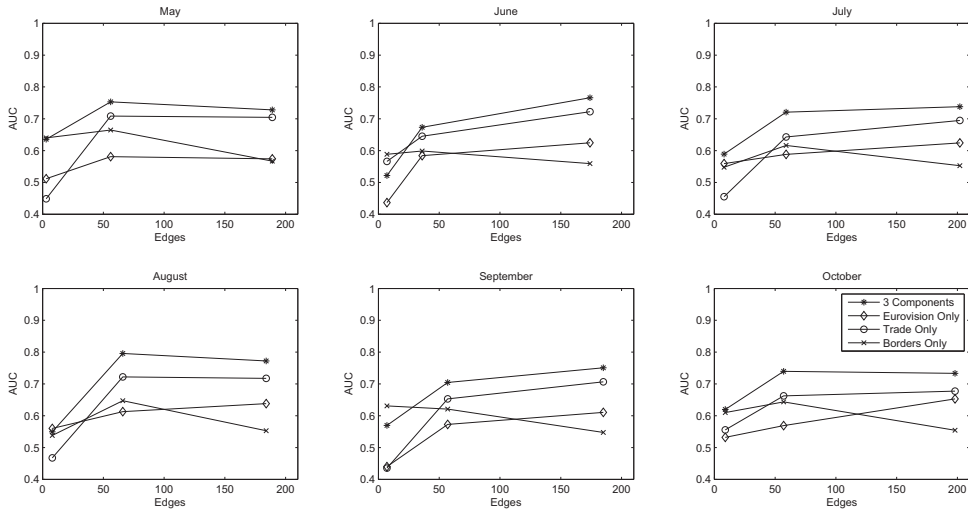


Fig. 10. Prediction AUC accuracies, under a 10-fold cross-validation scheme, for the ‘Coverage-similarity’ network for three different network densities over the six months period of study.

to name but a few. Developing methods to model and understand this kind of data is a pressing need in all those disciplines. This paper presented methods to deal with this kind of data, demonstrated in the context of news content patterns but directly applicable in other domains.

We have developed a statistical test for network similarity that allows us to conclude that we can define a stable relational structure in the content of EU outlets. We have then presented a predictive method based on GLMs, showing how this structure can be explained in terms of other complex relations, namely commercial, geographic and cultural relations between countries.

At the same time, we have presented what we believe to be the first large scale investigation of content patterns across 22 languages and 27 countries, involving the analysis of millions of news articles. We believe that this kind of studies will become important in the social sciences in the years to come.

ACKNOWLEDGEMENT

The authors want to thank Tijn De Bie, Marco Turchi, Omar Ali and Phil Naylor respectively for advice about data modelling, machine translation, data gathering and computer infrastructure support; the entire ‘Pattern Analysis and Intelligent Systems’ group at the University of Bristol for

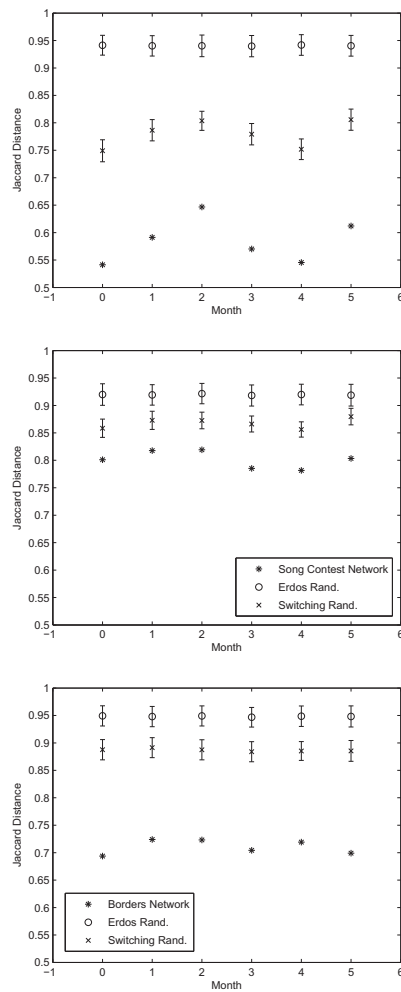


Fig. 7. Comparison of the 'Citations' network to (from top to bottom): 'Trade', 'Song Contest' and 'Borders' networks.

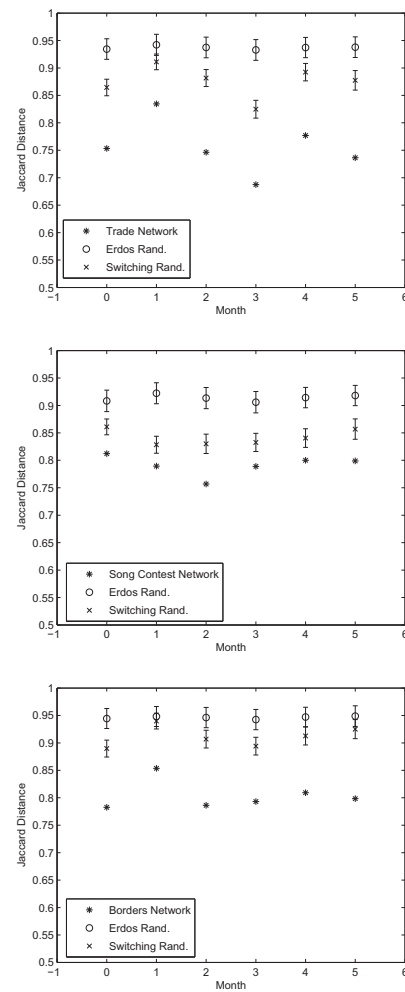


Fig. 8. Comparison of the 'Coverage-similarity' network to (from top to bottom): 'Trade', 'Song Contest' and 'Borders' networks.

discussions; and the Pascal2 network which supported the group activities. Ilias Flaounas is supported by Alexander S. Onassis Public Benefit Foundation, Nick Fyson is supported by the Bristol Centre for Complexity Sciences (EPSRC grant EP/5011214) and Nello Cristianini is supported by a Royal Society Wolfson Merit Award.

REFERENCES

- [1] I. N. Flaounas, M. Turchi, T. D. Bie, and N. Cristianini, "Inference and validation of networks," in *ECML/PKDD*, ser. LNCS, W. L. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, Eds., vol. 5781. Springer, 2009, pp. 344–358.
- [2] P. Koehn, H. Hoang, and et al., "Moses: Open source toolkit for statistical machine translation." in *Proceedings ACL '07, demonstration session.*, 2007.
- [3] M. Turchi, I. N. Flaounas, O. Ali, T. D. Bie, T. Snowsill, and N. Cristianini, "Found in translation," in *ECML/PKDD*, ser. LNCS, W. L. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, Eds., vol. 5782. Springer, 2009, pp. 746–749.
- [4] H. Kashima, Y. Yamaniishi, T. Kato, M. Sugiyama, and K. Tsuda, "Simultaneous Inference of Biological Networks of Multiple Species from Genome-wide Data and Evolutionary Information: A Semi-supervised Approach," *Bioinformatics*, vol. 25, pp. 2962–2968, 2009.
- [5] M. Pelillo, "Replicator equations, maximal cliques, and graph isomorphism," *Neural Computation*, vol. 11, no. 8, pp. 1933–1955, 1999.
- [6] H. Bunke, "Error correcting graph matching: on the influence of the underlying cost function," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 917–922, 1999.
- [7] M. Fernández and G. Valiente, "A graph distance metric combining maximum common subgraph and minimum common supergraph," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 753–758, 2001.
- [8] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, no. 26, pp. 290–297, 1959.
- [9] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *SIGCOMM*, 1999, pp. 251–262.
- [10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, p. 824, 2002.
- [11] J. Nelder and R. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.
- [12] P. McCullagh and J. Nelder, *Generalized Linear Models*. Chapman and Hall, London, 1989.
- [13] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers (Technical Report HPL-2003-4)," *HP Laboratories, Palo Alto, CA, USA*, 2003.
- [14] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. 4th Edition, Academic Press, New York, 2009.