J E Johnson

Vladimir Gorodetsky
Igor Kotenko
Victor Skormin (Eds.)

LNCS 3685

Computer Network Security

Third International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2005
St. Petersburg, Russia, September 2005, Proceedings

Springer

Networks, Markov Lie Monoids, and Generalized Entropy

Joseph E. Johnson

University of South Carolina,
Department of Physics,
Columbia, South Carolina 29208
jjohnson@sc.edu

Abstract. The continuous general linear group in n dimensions can be decomposed into two Lie groups: (1) an n(n-1) dimensional 'Markov type' Lie group that is defined by preserving the sum of the components of a vector, and (2) the n dimensional Abelian Lie group, A(n), of scaling transformations of the coordinates. With the restriction of the first Lie algebra parameters to non-negative values, one obtains exactly all Markov transformations in n dimensions that are continuously connected to the identity. In this work we show that every network, as defined by its C matrix, is in one to one correspondence to one element of the Markov monoid of the same dimensionality. It follows that any network matrix, C, is the generator of a continuous Markov transformation that can be interpreted as producing an irreversible flow among the nodes of the corresponding network.

1 Introduction

There is a broad spectrum of mathematical problems that involve the general theory of networks and the associated classification, optimization, and potentially even their dynamical evolution. By a network we mean a set of n nodes (points), some pairs of which are connected with a representative non-negative weight or strength of connection. Such a network can be represented by a connection (or connectivity, or adjancy) matrix C_{ij} whose off-diagonal elements give the non-negative 'strength' of the connection between nodes i and j in the network. Often that 'strength' or 'weight' is as simple as a '1' for a connection and a '0' otherwise. A network can be 'undirected' or 'directed' depending upon whether C_{ij} is symmetric or not thus indicating respectively a symmetric or asymmetrical connection between i and j. There may or may not exist a well defined 'metric distance' between the nodes or, equivalently, positions for the points in a metric space of some dimensionality, such as airports for airline networks, or substations for power or utility distribution networks. It is well known that the classification of different network topologies cannot be accomplished with just the eigenvalue spectra of the connectivity matrix as there are topologically different networks with as few as five nodes that have the same eigenvalue spectra. One root of the network problem is that although the network is exactly defined by the C matrix, there are n! different C matrices that correspond to the same topology because different C matrices result from different

V. Gorodetsky, I. Kotenko, and V. Skormin (Eds.): MMM-ACNS 2005, LNCS 3685, pp. 129–135, 2005.

nodal numbering orders. Most network problems become computationally intractable for more than a few hundred nodes.

We are interested in seeking useful metrics (functions of the C matrix) for the description of the topology of large networks such as sub-nets of the internet which might have from a hundred to a million nodes, and thus perhaps a trillion connection matrix values. To be useful, the metrics must be (a) rapidly computable, (b) intuitively meaningful, (c) should holistically summarize the underlying topology with a few variables, and (d) ideally would offer meaningful expansions that would provide increasing levels of topological detail. Mathematically, they should be (e) invariant under the permutation group on node numbering. We are specifically interested in the information flows of which originating node sends email or data to which destination node; and we are not initially interested in the underlying physical connectivity nor the path which the information traverses. Internet transmissions are extremely dynamic and thus to achieve some form of continuity, we envision constructing the C matrix with the summation of information transfers, over some time window δ, surrounding a time t for $C(t, \delta)$ thus representing the time evolution of the connection matrix. Given the number of connections, this problem resembles the representation of a physical gas in terms of thermo dynamical variables (such as temperature, volume, pressure, heat, and entropy). Generally, in such internet environments there is no meaningful location or position metric and distance is not usefully defined. As such pressure and volume, do not have a clear meaning without a distance function. Nor is it clear that what equilibrium is being approached, if any, and thus heat and temperature do not offer clear meanings. However, we suggest that the concept of both Shannon and generalized Renyi entropies [1, 2] can be well defined and summarize the order and disorder in the underlying topological structure.

Initially, how to define entropy on the connection matrix is not clear since both Shannon and Renyi entropies are defined as the log of the sum of the powers of the components of a vector, x_i, representing probabilities: $S = c \log_2 (b(\Sigma x_i^a))$ where $\Sigma x_i = 1$ and where a, b, and c are constants. As such these entropies represent the disorder in the underlying probability distribution. The disorder is a maximum with an even probability distribution and is a minimum when all the probability is in one cell with others having a value of zero. But the connection matrix columns or rows cannot be used as probability distributions since the diagonal of C is totally arbitrary. Even if we make some arbitrary choice of the diagonal values of C and normalize the columns, it is not clear what underlying topological 'disorder' we are measuring. In this work, we utilize our past work on the decomposition of the general linear group in order to answer both of these objections and to gain insight into how one might define these entropy metrics in useful ways that satisfy the requirements a-e above.

2 Background on Markov Lie Groups and Monoids

We had previously shown [3] that the transformations in the general linear group in n dimensions, that are continuously connected to the identity, can be decomposed into two Lie groups: (1) an n(n-1) dimensional 'Markov type' Lie group that is defined by preserving the sum of the components of a vector, and (2) the n dimensional Abelian Lie group, A(n), of scaling transformations of the coordinates. To construct the Markov type Lie group, consider the k,l matrix element of a matrix L^{ij} as a basis for n x n matrices, with off-diagonal elements, as $L^{ij}_{kl} = \delta^i_k \delta^j_l - \delta^j_k \delta^j_l$ with i =/= j. Thus the ij basis matrix has a '1' in position ij with a '-1' in position jj on the diagonal. These n(n-1) matrices form a basis for the Lie algebra of all transformations that preserve the sum of the components of vector. With this particular choice of basis, we then showed that by restricting the parameter space to non-negative values, λ^{ij} >=0, one obtains exactly all Markov transformations in n dimensions that were continuously connected to the identity as M = exp (s λ^{ij} L^{ij}) where we summarize over repeated indices and where s is a real parameter separated from λ^{ij} to parameterize the continuous evolution of the transformation. In other words λ^{ij} L^{ij} consists of non-negative coefficients in a linear combination of L^{ij} matrices. This non-negativity restriction on the parameter space removed the group inverses and resulted in a continuous Markov monoid, a group without an inverse, in n dimensions, MM(n). The basis elements for the MM algebra are a complete basis for n x n matrices that are defined by their off-diagonal terms. The n dimensional Abelian scaling Lie algebra can be defined by $L^{ii}_{kl} = \delta^i_k \delta^i_l$ thus consisting of a '1' on the i,i diagonal position. When exponentiated, A(s) = exp (s λ^{ii} L^{ii}), this simply multiplies that coordinate by e^s giving a scaling transformation. In what follows, we will show that all networks exactly correspond (one to one) to a combination of this Abelian transformation group and the Markov monoid transformations.

3 Connecting Markov Monoids to Network Metrics

The essence of this paper is the simple observation that (1) since the non-negative off diagonal elements of an n x n matrix exactly define a network (via C) and its topology with that node numbering, and (2) since a Markov monoid basis is complete in spanning all off-diagonal n ½ n matrices, then it follows that such networks are in one to one correspondence with the elements of the Markov monoids. Thus each connection matrix is the infinitesimal generator of a continuous Markov transformation and conversely. This observation connects networks and their topology with the Lie groups and algebras and Markov transformations in a well defined way. Since the Markov generators must have the diagonal elements set to the negative of the t sum of the other elements in that column, this requirement fixes the otherwise arbitrary diagonal of the connection matrix to that value also (sometimes referred to as the Lagrangian).

It now follows that this diagonal setting of C generates a Markov transformation by $M= e^{\lambda C}$. One recalls that the action of a Markov matrix on a vector of probabilities (an n-dimensional set of non-negative real values whose sum is unity), will map that vector again into such a vector (non-negative values with unit sum). The next observation is that by taking λ as infinitesimal, than one can write M = I + λC by ignoring order l2 and higher order infinitesimals. Here one sees that the bandwidth of the connection matrix between two nodes, now give that M matrix element as the relative transition rate between those two components of the vector. Thus it follows that given a probability distribution xi distributed over the n nodes of a network, then M gives the Markov transition (flow) rates of each probability from one node to another. Thus it follows that the connection matrix gives the infinitesimal transition rates between nodes with the bandwidth reflecting that exact topology.

Specifically, if the hypothetical probability vector is x_i=(1,0,0,0...0) then the first column of the M matrix will give the concentration of probability at the i^{th} node after that infinitesimal time period. Thus the first column of M is the probability distribution after an infinitesimal time of that part of the probability that began on node 1 and likewise for all other nodes thus giving a probability interpretation to each of the columns of M. Thus each column of M can be treated as a probability distribution associated with the topology connected to that associated node and supporting an associated entropy function that reflects the inherent disorder (or order) after a flow λ. Thus the columns of M support a meaningful definition of Renyi entropies which in turn reflect the Markov transformation to disorder of the topology near the node for that column. Thus this Renyi entropy on this column can be said to summarize the disorder of the topology of the connections to that node. It follows that the spectra of all nodes reflects in some sense the disorder of the entire network. When sorted in descending order, it represents a spectral curve independent of nodal ordering and thus independent of the permutations on nodal numbering. That spectral curve can be summarized by the total value for the entropy of all columns (since entropy is additive and the column values are totally independent).. The structure of the spectra can also be summarized by the entropy of the entropies in the spectra thus giving a second variable summarizing the entire topology.

If the connection matrix is symmetric then the graph (network) is said to be undirected, but if there is some asymmetry, then the graph is at least partially directed where the flow from i to j is less or greater than the converse flow. If the connection matrix is not symmetrized then one can capture this asymmetry by resetting the diagonal values of C to be equal to the negative of all other row values in that row. Then upon expansion of $M = I + \lambda C$, the rows are automatically normalized probabilities that in turn support entropy functions for each row. These row entropy values form a spectrum which could be sorted by the same nodal values (in order) that is used to order the column values. This will result in a different spectral curve that is not necessarily in non-decreasing order for the row entropies. One also can compute the total row entropy and the entropy if these row entropies as we have done from columns. If two columns have the same entropy then one can sometimes partially remove this degeneracy by the values of the associated row entropies.

Thus we suggest that the column and row spectral entropy curves, and the column and row total entropy and entropy of entropy values, distil essential disorder and order from the network topology – from n^2 values down to 2n (spectral) values, and finally to 4 values for the entire network – constitute a set of entropy metrics for the network, all of which are independent of the nodal ordering (numbering) in the network and thus indicative of the underlying topology. This analysis is expansive in two ways: (1) These two spectra and four values can be computed to higher order in λ thus including higher orders of the C matrix approximation for M and thereby incorporating connections of connections into the metric values. It is with higher powers of C via larger values of λ that we unfold more complex aspects of the network topology. (2) One can also compute these metric values for each of the Renyi entropy values. Work by V. Gudkov [4] has found that the order of the Renyi entropy is equivalent to the Hausdorf dimensionality equation. This opens the possibility that higher order entropy reveals connections of a 'higher dimensionality' in the network structure [4, 5].

4 Expansion of Second Order Renyi Entropy as a Taylor Series

Let us assume that C is symmetric (an undirected graph) thus $C = C^T$. If one considers the expansion of a vector of probabilities from state $\lambda=0$, $|x(0)>$, to another vector at a later state λ, $|x(\lambda)>$ by the continuous Markov transformation $M = e^{\lambda C}$ then $|x(\lambda)> = e^{\lambda C}|x(0)>$ and thus the entropy is given by:

$$S = \log_2(n\Sigma x_i^2) = \log_2(n<x(\lambda)|x(\lambda)>) = \log_2(n<x(0)|(e^{\lambda C})^T(e^{\lambda C})|x(0)>)$$

or rearranging and defining R we get:

$$R(\lambda) = 2^S/n = <x(0)|\, e^{2\lambda C}\, |x(0)>) \quad \text{since } C = C^T$$

and then expanding the exponential we get:

$$R(\lambda) = <x(0)|\,(I + 2\lambda C + (2\lambda C)/2! + \ldots)\,|x(0)>$$

Thus this power of the second order Renyi entropy consists of two times the diagonal values of the powers of the connection matrix, plus the unit matrix as shown. From this one can see that as λ becomes larger and larger, one must take more and more of the topology connections into consideration. This in fact gives a hierarchical expansion of this entropy that gradually 'explores and includes' higher and higher order connectivity. If the row and column entropies are computed to include these higher orders, then they will begin to take into account more complex aspects of the networks interconnectedness. When there is asymmetry a similar equation can be obtained.

5 General Diagonal Values and Eigenvalues

The previous results can be generalized to include totally general diagonal values for C, by utilizing the diagonal transformations available in the n-parameter Abelian scaling group. This group simply multiplies any node value by a scaling factor via $M= e^{\lambda C}$. There is a natural interpretation to the actions of this group in terms of network probability flows as introducing a source or sink of probability at the node which is acted upon. That action removes the conservation of probability that was maintained by the Markov monoid, but since such flow was simply used to encapsulate the topological structure of the network, we can accept this lack of conservation. Thus one can add to any diagonal of C, any positive or negative value representing the scaling value of that coordinate and one will still have a valid network as all off diagonal values of C are unchanged and the M matrix will still give the indicated flows. This allows one to see the previous arbitrary allocations of '1' or '0' of the C diagonals in a new light, especially for the eigenvalue computations.

When C is diagonalized, with the values leading to the Markov transformations, or to the more general values of the diagonals of the last paragraph, one automatically gets a diagonalization of the M matrix. The interpretation of the eigenvectors is now totally obvious as those linear combinations of nodal flows that give a single eigenvalue (decrease when the transformation is Markov) of the associated probability, for that eigenvector. This follows from the fact that all Markov eigenvalues are less than one except the one value for equilibrium which has eigenvalue unity for equilibrium. That means that each of these eigenvalues of C reflect the decreasing exponential rates of decrease of the associated eigenvector as the system approaches equilibrium as λ approaches infinity in $M= e^{\lambda C}$. This insight allows us to see that all of the Renyi entropy values are increasing as the system approaches equilibrium, which is normally the state of all nodes having the same value of this hypothetical probability. The use here of this 'artificial flow of probability under M' provides us with more than just a method of encapsulating the topology with generalized entropy values, it also gives an intuitive model for the eigenvectors and eigenvalues for C and sheds light on the graph isomerism problem (different topologies having the same eigenvalue spectra).

6 Conclusion. Potential Applications to Large Internet Networks

Based upon the arguments above, we suggest that for real networks such as the internet, that the appropriate connection matrix be formed, from source and destination information transfers, where both asymmetry and levels of connection are to be maintained in the C(t) matrix values during that window of time about that time instant. Specifically, this means that if a connection is made multiple times in that time interval, then that C element should reflect the appropriate weight of connectivity as this adds substantial value to the entropy functions. We then suggest that at each instant, the column and row entropy spectra be computed along with the total row and column entropy and entropy of entropies and that this be done for lower order Renyi entropies as well as lower order values in the expansion of the Markov parameter λ that includes higher order connectivity of the topology. We are currently performing tests to see how effective these entropy metrics are in detecting abnormal changes in topologies that could be associated with attacks, intrusions, malicious processes, and system failures. We are performing these experiments on both mathematical simulations of networks with changing topologies in know ways, and also on real network data both in raw forms and in forms simulated from raw data. The objective is to see if these metrics can be useful in the practical sense of monitoring sections of the internet and other computer networks. In addition to the two values of total entropy and entropy of entropy that summarize the column (or row) spectral distribution, we are looking at other natural expansions of this function in terms of functions or orthogonal polynomials that summarize the general behavior in useful ways thus providing other summary metric variables for the entropy spectra.

Acknowledgements

The author benefited from extensive collaborations and conversations with Dr. Vladimir Gudkov.

References

1. Renyi, A.: Probability Theory, North-Holland Series in Applied Mathematics and Mechanics, North-Holland Pub. Co (1970) 670 pages.
2. Renyi, A.: Selected Papers of Alfred Renyi, Akademia Kiado, Buadapest, Vol. 2 of 3 volumes (1976)

3. Johnson, J.E.: Markov-type Lie Groups in GL(n, R). Journal of Mathematical Physics 26 (1985) 252–257
4. Gudkov, V., Johnson, J.E.: Network as a complex system: information flow analysis, arXiv:lin.CD/0110008v1 (2001) 10 pages
5. Gudkov, V., Johnson, J.E.: Chapter 1: Multidimensional network monitoring for intrusion detection, arXiv: cs.CR/020620v1 (2002) 12 pages