Mathematical Foundations of Networks Supporting Cluster Identification

Joseph E. Johnson and John William Campbell
Physics Department, University of South Carolina, 730 Main St. 29208, Columbia SC, U.S.A.

Keywords: Cluster Identification, Network Series Expansions, Renyi Entropy Spectra, Markov Type Lie Group, Continuous Groups, Lie Group.

Abstract: The author proved that the continuous general linear (Lie) group in n dimensions can be decomposed into (a) a Markov type Lie group (MTLG) preserving the sum of the components of a vector, and (b) an Abelian Lie scaling group that scales each of the components. For a specific Lie basis, the MTLG generated all continuous Markov transformations (a Lie Markov Monoid LMM) and in subsequently published work, proved that every possible network as defined by an n x n connection matrix C_{ij} of non-negative off-diagonal real numbers was isomorphic to the set of LMM. As this defined the diagonal of C, it supported full eigenvalue analysis of the generated Markov Matrix as well as support of Renyi entropies whose spectra ordered the nodes and make comparison of networks now possible. Our new research provides (a) a method of expanding a network topology in different orders of Renyi entropies, (b) the construction of a meta-network of all possible networks of use in network classification, (c) the use of eigenvector analysis of the LMM generated by a network C to provide an agnostic methodology for identifying clusters and (d) an a methodology for identifying clusters in general numeric database tables.

1 INTRODUCTION AND PREVIOUS RESEARCH

Prior work by the first author established a general mathematical foundation for the theory of networks that is extended by the current research on network series expansions and cluster identifications. It is common knowledge that vast domains of knowledge can be expressed in the form of networks. Furthermore our understanding and classification systems of the world and even of language itself depend upon the concept of clustering within those networks. We first review the foundations of the underlying mathematics and that of networks along with the first author's previous results (Johnson 2005) and (Johnson 2006) in order to frame our current joint results.

1.1 Background on Networks and Cluster Analysis

A network is here defined as a set of 'n' points called nodes and numbered 1, 2, …n along with a set of connections among those nodes given by real non-negative numbers $C_{ij.}$ These values are to represent the 'strength of connection' between nodes i and j and are specified by a square n * n matrix C, normally called the connection, adjacency, connectivity, or network matrix. One example of C is to assign a number 1, 2, …to each member of a group and then define C_{ij} to be the number of emails which each person i sends to another person j per month. Obviously C is also a function of time and also is not symmetric since every email i to j does not necessarily have another one that goes from j to i.. It is also normally the case that the vast majority of the C matrix values are '0' in value as a given node will only connect to a few hundred or few thousands of other nodes. For example a person usually has less than a thousand contacts for phone or email out of the 7 billion people on earth. Thus C is called a 'sparse matrix' predominantly consisting of zeroes. The connections are not allowed to be negative because one cannot have less than a '0' (no) connection between two nodes. The diagonal terms are not defined for C because one cannot give a "strength" to a connection of a thing with itself. Thus it is important to realize the C diagonal is not equal to zero but rather is not defined at all. The connection from i to j is normally independent of the connection from j to i thus making the C matrix

E. Johnson J. and William Campbell J..
Mathematical Foundations of Networks Supporting Cluster Identification.
DOI: 10.5220/0005086902770285
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2014), pages 277-285
ISBN: 978-989-758-048-2

different from its transpose (C_{ij} != C_{ji}) and thus is asymmetric as a matrix. A special set of networks have only the values C_{ij} = 1 or 0 and are called 'graphs' so they either have a connection or not and can be directed (not equal to their transpose) or not. Although this is a special case of the networks that we study, they are highly degenerate and the richness of the different real number values of C is lost. Networks normally have certain nodes that have more connections than most nodes in the network, and are called 'hubs'. Groups of nodes often are much more interconnected with each other than with other nodes and form what are called "clusters". Cluster analysis is an essential component of knowledge, and even language as we give common entities and concepts a group name if their properties are similar and thus cluster. If every member of a group of nodes has a non-zero connection with every other member then this sub-network is called a "clique" and is represented by a dense set of connections. Other common networks are ""tree", "star", "ring", "random", and "scale-free" networks. Networks have the unusual property that one can take any subgroup of the nodes and consider that "sub-net" to be a node by collapsing that subnet to a single node. For example all computers that belong to a given corporation could be thought of as a single node as represented by the company and thus ignoring all internal traffic. Conversely all nodes on the internet could be grouped into the "outside" and only those nodes within the corporation might be considered in constructing C. In conclusion, a network among n nodes is defined as a n x n matrix C_{ij} consisting of non-negative real values and with undefined diagonal elements and thus is defined by n x (n-1) independent non-negative values.

1.2 The Fundamental Problems with Network and Cluster Analysis

At first glance, it would appear easy to classify networks and clusters the way we classify matrices in mathematics but this is far from the case: (a) The mandated non-negativity of C values disallows the full range of real numbers as a constraint. (b) The absence of definition for the diagonal terms leaves the essential definition of the matrix undefined and without the ability to perform the primary analysis of eigenvalue and eigenvector determination, (c) The lack of a unique and well defined ordering for the nodes and thus for a unique definition of C means that there are n ! different C matrices that equivalently describe each network and thus we cannot even tell if two networks are the same. (d) The lack of a unique, non-arbitrary definition of the concept of cluster as there are over 100 common algorithms and definitions of what constitutes a cluster. (e) The very large sizes of the C in the practical world can defeat real computations as n can be perhaps seven billion squared to describe just one kind of network of humans. (f) The actual number of nodes often changes over time or by the nature of the problem being studied, by a splitting or merging of sets of nodes. Thus new nodes can spring into existence and others can disappear causing the matrix to not even have the same size from moment to moment. (g) The distance between two networks has no natural meaning and thus one lacks a metric for the space of networks. Specifically this means that one cannot define the rate of change of a network over time dC/dt and this defeats a dynamical theory of networks. (h) Finally, with most complex systems in the sciences, there are means for "expanding" the system in a series of sequentially less important terms such as Fourier expansions of sound waves, binomial and Taylor expansions of functions and multipole expansions of mass and charge distributions. Such expansions could capture the most important or dominant aspects of the topology would be invaluable in beginning a classification system or for comparing two networks as well as simplifying our network descriptions but there is no such system for networks. (i) There is no "intuitive or physical model" for networks that can guide us in deeper understanding and guide our intuition. (j). We also lack any definitions for invariants or conserved quantities such as energy, momentum, mass, charge, and angular momentum. Nor do we even have metrics for concepts such as "temperature", "entropy", or "information content". (k) Finally, there is no well-defined concept of what is "optimal" for a given network (given some 'purpose") and thus one cannot measure "how far from optimal" a given network is, or how rapidly it is approaching such an optimal configuration. This list of problems is not even exhaustive.

1.3 Background in Continuous (Lie) Markov Transformations

Markov transformations are linear (matrix) transformations that, when acting on a vector of non-negative values (positive or zero components), preserve the sum of that vectors components (i.e. the sum of the components is invariant) and give a new vector that also has non-negative components. Thus

while a rotation leaves the sum of the squares of a vectors components invariant and describes the motion on a circle or n-dimensional sphere, a Markov transformation leaves the linear sum invariant and describes the motion on a straight line, or plane or generally a hyperplane that is perpendicular to the vector (1,1,1,…..) and where all vectors, both before and after the transformation are in the positive hyperquadrant. Markov transformations describe diffusion and increasing disorder such as the dispersion of ink into clear water or dirt in one's home. They are the transformations that describe the irreversibility of time, increasing entropy (disorder) and the gradual loss of organized energy into heat (random energy) and thus the second law of thermodynamics (and even the loss of information in systems). Because Markov transformations do not have an inverse, they were never studied from the point of view of group theory, because mathematical groups all have inverse transformations (along with closure, an identity, and associativity). It can be shown that all Markov transformations are square matrices that consist of non-negative (positive or zero) numbers where each column sums to unity (one). Another type of Markov matrix has the row values sum to unity. Essentially all studies of Markov transformations are for discrete and not continuous transformations. It is the continuous Markov transformation that will be central to our work related to networks.

A mathematical group is a set of objects (say A, B, C, …) and a multiply operation (say *) that has (a) closure into another member of the set), (b) is transitive i.e. the ordering of the operation among three elements does not matter, (c) has an identity transformation leaving another element unchanged, and (d) for every element the group has an inverse that reverses the action of the first. One simple example is the set of the identity and the reflection R in a mirror. Another example is the set of four rotations of a square that leave it invariant (by 0, 90, 180, and 270 degrees). Then one can consider the group of rotations about an axis or the translations on a straight line as examples of continuous (Lie) transformations of rotation both of which have an infinite number of elements. In the 1890s Sophius Lie invented a way to study all of these by studying the associated infinitesimal transformation where he showed that an exponentiation of the infinitesimal transformation gives the original transformation. This means that we can study a single transformation L rather than the infinite number of rotations. For rotations in three dimensions, there is a set of three such transformations: Lx, Ly, and Lz for rotations about each axis. Thus one only has three objects that are needed to study all of the three-fold infinity of rotations in three dimensions. The resulting set of L matrices is called the Lie algebra for that Lie group, R, which is generated by exponentiation. This group is called the rotation group R3 or the Orthogonal group O(3).

1.4 Decomposition of the Continuous Linear Transformation Group

The general linear group of all continuous transformations in n dimensions is represented by an n x n (invertible) matrix of real numbers. Such transformations include rotations, translations, and the Lorentz transformations of the theory of relativity as well as all the unitary transformations in quantum theory. Transformations allow us to study symmetry such as rotational symmetry or other invariance. Since we wish to generate all continuous linear transformations, we will need all possible infinitesimal generating matrices which are easily listed as having a '1' in the i,j position and a '0' in all other positions. There are (as might be expected) n^2 such matrices since we can put the '1' in any of the n^2 positions. Those matrices with a "1" in the i, j position form the n^2 elements of the general linear group. However, it was discovered by the author (Johnson 1985) that the general linear group can be decomposed into two separate Lie groups as follows: (a) Consider the generator (Lie algebra) element which has a 1 at the ii position and a 0 at every other position. If we exponentiate that matrix then this is obviously e^a at one diagonal position, 1 at other diagonal positions, and zeroes everywhere off the diagonal. These transformations multiply that one axis by e^a and multiply all the rest by '1' thus leaving them unchanged so it just makes that one axis longer or shorter by that factor. We call these scaling transformations and the group is called Abelian because every transformation commutes with all the other elements in the algebra. We next identify the Markov Type Lie Group (MTLG). Consider the off-diagonal algebra (generators) and rather than using just a '1' at each off diagonal position, let us form an element by placing a '-1' on the corresponding diagonal of that column. This makes the sum of the elements in each column of the generator equal to zero with a "1" off the diagonal and a "-1" on the diagonal in the same column. Every other value is "0". Formally this defines the m,n matrix element. There are obviously n^2-n such L matrices corresponding to every position off the

diagonal. The first author showed that the exponentiation of these L matrices always generates a Markov type matrix and conversely all continuous Markov type matrices (transformations are so generated. We call this Lie group and its associated Lie algebra a 'Markov Type' Lie group or algebra because it preserves the sum of the components of a vector. However it does not preserve the positive definiteness of the components of the vector because M can also take one to negative values of the coordinates. We now restrict the MTLG to give only physically acceptable transformations resulting in a Markov Monoid (MM). One can easily verify that if the parameters that multiply the L generators are all non-negative, then one only gets a transformation that takes one from a vector with only non-negative values to another vector with non-negative values. This is essential if the vector components are to represent probabilities (or numbers of objects). In that process however, one gives up the inverse of the transformation and we end up with a group without an inverse which is called a Lie 'monoid'. Now with this restriction to non-negative values of the L multipliers, we always get a Markov matrix: One notices that the sum in each column is '1' and that all elements are positive. This result was highly significant because it tightly connected the theory of Lie groups and Lie algebras with the theory of Markov transformations allowing the theorems and insights in that powerful domain of Mathematics to be utilized in the other domain: the theory of all continuous Markov transformations.

1.5 Networks Are 1 to 1 (Isomorphic) with the Lie Algebra for Markov Transformations

The author (Johnson 2005) subsequently proved that every network is a Markov monoid (MM) and conversely. Recalling that any network is an off diagonal set of non-negative numbers, it now follows immediately that we can multiply the appropriate MM generator by the value in the off diagonal value of a given network, and end up with a Markov monoid matrix that will generate a valid Markov transformation. This is a consequence of the fact that each diagonal is automatically defined as the negative sum of off-diagonal elements in that respective column. Thus any network C gives exactly one MM Lie generator for a continuous Markov transformation and conversely any MM generator defines, via its off-diagonal elements, a network with that C matrix with exp(aC). This important result now connects the study of the complete topology of all networks to the study of the equivalent MM and its associated unique Markov transformation. The collective power of three branches of mathematics, Lie algebras & groups, Markov transformations, and Networks, are now fully integrated allowing us to use the power of each domain to study the other domains. This result also has an immediate positive consequence, namely that the diagonal of the C matrix is exactly defined and is unambiguous as a MM where each diagonal element is the negative of the sum of the off diagonal terms in that corresponding column. This puts network theory on a firm unambiguous mathematical footing and every possible network defines a continuous Markov transformation in that number of dimensions as defined by the associated Lie and Markov monoid.

These results provide solutions to each of the core network problems as follows. The first important consequence is that since the C matrix now has its diagonal determined and is unambiguous, that all eigenvector and eigenvalue analysis is well defined. With some thought, one can show that the associated eigenvalues are all '0' or negative with the '0' value being associated with the equilibrium eigenvalue, and all the other (negative) eigenvalues being associated with flows of an associated diffusion rate that is exponentially deceasing for the corresponding eigenvector and representing an approach to equilibrium for the vector upon which it acts. There are also cases (since the resulting C matrix may not be 'normal' (as required in order to have real eigenvalues), where there can be complex eigenvalues and in this case this eigenvalue gives the angular velocity of a circular flow of conserved entity under the Markov transformation while the real component provides the decay of that cycle to zero (equilibrium). This is very analogous to the physical system of coupled harmonic oscillators with overdamped, critically damped and underdamped solutions. In fact, in spite of the fact that the network matrix is a static system, it can be modeled by a dynamical evolution of the approach to equilibrium of the diverse combinations of nodes that constitute each eigenvalue and which approaches zero if time were to evolve the associated eigenvalue. This is an example of where we can use dynamical evolutions of the associated 'time' parameter to inform us of the structure of the C matrix using the time evolution of the MM.

Renyi entropies of order two can be defined on each column in the resulting Markov matrix M thus providing another set of critical metrics for the topology. This second equally or perhaps more

important consequence is that since the resulting Markov matrix has columns that are non-negative and sum to unity, each column can be interpreted as a probability distribution. Thus it follows that each column can support a well-defined concept of entropy (either Shannon or Renyi') on each column. This entropy value measures the order or disorder of the incoming (columns) or outgoing (rows) flows of the conserved substance such as probability to the node in question as per the model which we described above. Thus each column (and each row separately) has a numerical value that can be used to either partially or totally distinguish them, and which can be used to uniquely number the nodes! By sorting the Renyi entropy values in order, we obtain an entropy "spectral curve" that is highly descriptive of the topology. No two topologies can be identical unless the entropy spectral curves are identical and thus we can take the distance between the Renyi entropy curves as a measure of the distance between the two topologies (computed as sum/integral of the curve differences squared). In the previous work we only considered the use of the two second order Renyi entropy spectral curves where one was computed for the columns and one computed for the rows. The distance metric between network A and B was defined as the sum of the column and the row distances between the two networks.

1.6 The Algorithm for Network Analysis with Entropy Metrics

It is easier here to bypass other details of the technical foundation and give the exact prescription from the following algorithmic steps (Johnson 2012). Set the diagonal terms C of a connection matrix to be equal to the negative of the sum all elements in that respective column (and then later redo all this for the rows instead of the columns to achieve a second type of Markov transformation). Then divide every element of the matrix by the negative of the trace *n thus 'normalizing' the matrix to have a trace of '-n'. It can be shown that this matrix is the infinitesimal generator of a continuous Markov transformation since it is a linear combination of the Markov monoid Lie generators. Compute the associated Markov matrix as exp(aC) using any number of terms and one will always get, in any order, a Markov matrix where all matrix elements are non-negative and the sum of all elements in any column is '1'. The number of expansion terms used represents the degrees of separation thus incorporated and this is an important consideration in informing the entropy functions that are computed in terms of the M matrix of the number of degrees of separation to be considered. Since each column has only non-negative elements and each sums to unity, it now follows that these elements can be interpreted as probabilities and thus support a definition of entropy which is defined as the negative of the log of sum of the squares of the elements of that column. As this S_j is defined for each column (node), we may sort the nodes in order of these values. For real values there is rarely degeneracy, but if there are two or more equal values, then a similar procedure, when performed on the rows, will usually distinguish the sort order and if not, one uses the higher Renyi orders. These sorted values provide an "entropy spectra" for the columns which can be plotted as a curve, and likewise one obtains another entropy spectral curve for the rows. These two curves for the row and column entropy spectra for each order of the Renyi entropy are specific to the network topology, represent the incoming and outgoing order/disorder of connectivity. The isomorphism of a network C to the Lie Monoid generator L occurs because (a) both C and L have all possible non-negative off-diagonal elements of a square matrix of any size, and (b) C has an arbitrary diagonal while the diagonal of L is defined as having a diagonal consisting of the negative of the sum of all non-diagonal elements of the corresponding column (or row) thus providing that definition for the diagonal for the network matrix C. It is precisely those Markov type generators that have a negative off-diagonal term that are not pertinent to the concept of a network and are exterior to our investigation. The MM Lie generators provide a definition of the diagonal thus allowing (a) a well-defined matrix whose eigenvalue and eigenvector structure can be studied, and (b) thereby providing a dynamic model of exponentially decreasing flows of all eigenvalues toward zero except for the eigenvalue of "0" which represents final equilibrium with maximum entropy. Secondly, that same MM generator always generates a family of Markov transformations whose column (or row) sums gives non-negative values that always sum to unity, and thus can be treated as probabilities.

2 NETWORK CLASSIFICAITONS, EXPANSIONS, AND CLUSTRING

The previous research enabled one (a) to determine a

unique diagonal for the C matrix making it a member of the Lie Monoid that created a family of Markov matrices; (b) thus allowing one to compute the unique associated eigenvectors and eigenvalues for the resulting Markov matrix and thus for the network topology; and (c) thus in turn to create a hypothetical physical model of dynamic flows among the nodes that modeled flows where each eigenvector (as a linear combination of nodes) would describe a flow (of an imaginary dispersing fluid or substance) toward equilibrium at the rate of the associated eigenvalue. The flow rate occurs among the nodes at rates proportional to the connection strength between those nodes. Consequently any network could be studied from the point of view of the unique associated Markov transformation and the associated physical model of approach to equilibrium of a dispersing system of a conserved substance (since the Markov transformation preserves the sum of the components of the vector upon which it acts) . Thus one now has the powerful tools of eigenvalue/eigenvector analysis to use in the study of the network topology. This work next enabled one to compute the (second order) Renyi entropies of the n columns and the n rows of the associated Markov transformation that which, when sorted, provided an entropy spectra curve (d) that ordered the nodes and allowed direct comparison between two networks to see if they were the same and (e) allowed one to define a distance metric between two topologies (as the length of the vector which is the distance between the two respective Renyi entropy vectors. This also then allowed one (f) to compute the distance between a network at one time and at a later time and thus compute the rate of change of a networks topology. But our past work did not provide (g) (a means of expanding a network in a series of terms which would reflect smaller and smaller aspects of the topology, (h) any framework that could support a classification of network topologies, or (j) any insight into the very complex structure of clusters in networks and the clusters within those clusters etc. We now have foundations laid in each of these areas (Campbell 2014).

2.1 Expansion of Networks using Higher Order Renyi Entropies

Networks can be uniquely identified by ordering the nodes using the Renyi entropies as previously discussed and these entropy curves based upon the MM matrices must be identical if the generating topologies are identical. First of all this essentially solves the problem of distinguishing and ordering the nodes using the sorted column (and row) entropies. But there are only 2n of the second order Renyi entropies as computed for each row and each column of the network generated Markov matrix while there are n^2 elements of the matrix. So even if the second order Renyi entropies are sufficient to provide a unique ordering of the nodes, they are not functionally sufficiently rich to carry all the network information. However one notes that the successive higher orders of Renyi entropies for each column and each row are functionally independent as each is proportional to the log of each sequentially higher powers of the components of the components of each column and row. Since the sum of the successive powers 2, 3, 4, …m are linearly independent then one only needs to utilize a sufficient number of powers (and thus orders of Renyi entropy) to determine the Markov matrix elements and thus the topology. Thus $2m = n^2 – n$. It is easy to see from the definition of the entropy that an ordering of nodes using one order of Renyi entropy cannot conflict with the ordering of another order of Renyi entropy. But aside from the functional independence of the sums of sequentially higher powers, the next most important realization is that for a given column or row, the sum of each higher power results in a value less than that of the lower power since all values are less than unity. It then follows that each higher order Renyi entropy is lower than its predecessor and that the distance between each successive curve (as previously defined) is smaller and smaller. Thus the set of curves that are the differences between the n and the n+1Renyi entropy is increasingly smaller and thus when they are taken together, they both represent the topology as a decreasing expansion of functional differences. Furthermore they are functionally complete to (only in principle) determine the entire original topology of $n^2 – n$ values.

A metric can now be defined for the distance between two topologies as follows. The closer these two entropy spectral curves are to each other, then the more similar the values of the incoming and outgoing entropy probability vectors. Thus we can usefully define the ‘distance between two topologies as the distance between these entropy spectral curves’ of the same Renyi entropy order as indicative of the ‘distance between the two topologies’. If the networks have the same number of nodes, then one can just take the sum of the squares of the differences of the two corresponding entropies. But in many cases one must compare a topology with another where an additional node (or

one missing node) prevents this direct comparison. In that case one takes the difference between the (boundary-normalized) smoothed sorted entropy curves and integrates the square of the distance between the curves and then takes its positive square root (like a scalar product in a Hilbert space). This essentially treats the Renyi entropy values for the columns (or rows) as a vector of n ordered components for each network and then takes the distance between the two topologies as the magnitude of the difference vector (the square root of the integral of the differences (or if they have the same number of nodes, the sum of the entropy differences). These differences can be summed over each order of Renyi entropy for both the rows and columns to define a total distance between the topologies. Now that we have a metric for the distance between two topologies, we can take the derivative over time of that change and identify aberrant changes or differences. We have previously applied this to a study of system attacks on networks (using only the second order entropy) and for the identified aberrant changes. This method was successful in identifying network attacks not see by other software and also identified abnormal usages in large university network flows.

2.2 A Framework for Potential Network Classification

We ask the reader to now imagine an exceedingly large mathematical network where each node is itself a network. This mega-network is then to be defined by a connection matrix that consists of the exponentiation of the negative of the total summed distance between the successive Renyi entropies for the columns and where the transpose terms are defined the similar term using the rows. Then two networks are closer when these distances are smaller and since we exponentiate the negative of this positive distance it follows that when the distance becomes large, the connection becomes small as we would desire. This "MetaNetwork" is one of the largest entities in all of mathematics as every possible network (n*(n-1) set of non-negative reals) will constitute a node. The virtue of such a MetaNetwork is that although the dimensionality is very, very large and certainly the nodes cannot be positioned in a 3 or even a finite dimensional space, one can use special networks (trees, rings, clusters, scale-free networks, etc.) as reference points or better yet with sequences of them as axes with which other topologies can be referenced and thus positioned in this space Even to create this C, one must truncate the number of networks so that the diagonals can be determined. We will present our limited results for meta-networks containing some finite number of networks as nodes. In conclusion, whereas one cannot determine the "coordinates" of a network in this mega-network, one can find the distances from a network in question to a large number of "reference networks" (trees, rings, clusters, scale-free, and random networks).

2.3 Eigenvectors Determine Network Clusters

There are well over 100 different methods for mathematically identifying clusters. What one chooses to define as "similar" for clustering can vary greatly and there seems to be no natural definition. Intuitively one understands the basic concept of a cluster as a group of items in a set that have "very similar properties" or in a network as a subnet that is "highly connected". It is the "cluster" that in many ways is the foundation of our language, concepts of abstractions, classifications, and intelligent reasoning and thus of the greatest possible importance. If we consider the eigenvectors of a network, they represent those combinations of nodes, (with some weighting vector that is normalized to unity), that, like the normal nodes of a set of coupled harmonic oscillators, will approach equilibrium at the unique rate of the associated eigenvalue. The nodes, weighted as per the defining linear combination for the eigenvector, behave in the model as one entity and transfer the imaginary substance in the vector being acted upon by the MM, among themselves. It follows that it is the eigenvectors that constitute clusters with nodes participating in a given cluster, in proportion to the weights, that provide a neutral definition of a cluster. The Markov matrix generated by a network has been shown by the authors to have eigenvectors that identify not only clusters but also the complex structure of such clustering within clustering. The eigenvalues and eigenvectors of the M matrix provide an intuitive model for networks in the following way. The continuous one-parameter transformation generated by the network monoid serves as a dynamic model for any network as a combination of conserved flows among the components of a vector upon which it acts. Specifically the eigenvectors become those linear combinations of nodes that have unique associated eigenvalues representing the rate of flow toward equilibrium, or when complex, the angular frequency of flow cycles in the approach to

equilibrium. Clusters in networks can be identified by the magnitude of the component network nodes within each eigenvector. Consequently one can see the multiple levels and complex structure of what constitutes "clustering". We have demonstrated this first with the direct construction of networks with clusters and even clusters within clusters. In each case the eigenvector identified such structures. The eigenvectors of the Markov matrix computed to different levels of connectivity also reveals additional structures. Our definition based upon the M eigenvectors is general and is agnostic to any arbitrary or additional assumption. We are not able to "prove" that these eigenvalues identify the clusters because there is no fixed formula that defines a cluster. But based upon the rate of flows within the eigenvector being maximal and contained in that eigenvector, this forms what our intuition would suggest is the most neutral definition of a cluster.

2.4 The "Properties" of "Entities" as Defining a Network with Cluster Identification

In parallel research, we constructed a new kind of network as follows. Imagine that one has a set of entities (chemical elements with numeric properties, economic profiles of companies, or people with numerical properties). Assume that all the properties of all entries are numerical values which measure the extent of that property that the entity possesses. We can define each entity to be a node and form a network among the entities as follows: Normalize each property column by transforming the values to the number of standard deviations for that column away from its mean value which we rescale to zero. This then removes the units in each column and puts the columns in an equivalent form of standard deviations. (If the values cover several powers of ten, then one would use the more reasonable value of the log of the values of the properties). Then one can form the function that is the sum of the squares of the differences of the respective values of each property and then exponentiate the negative of this value. Notice that this function is very close to zero when the entities have very similar properties (in terms of the standard deviations from the norm). Thus the function will be the largest when the entities have extremely similar properties and thus are very much alike. This is somewhat like computing the probability that entity x is the same as entity y. By creating this C matrix for all of the entities, one creates a network among the entities that shows strong connections between highly similar entities. The study of the clustering in such a network via a study of the associated MM generated Matrix and its associated eigenvectors show unique and well defined clustering of the entities in terms of the properties listed. . We have performed this both for the periodic table of elements and found reasonable clustering of physical properties of the elements and also for the Leontief IO model economic sectors of the U.S. economy using the use and make matrices at the 100 level of disaggregation.

3 CONCLUSIONS

Our new results first provide a powerful tool for the expansion of a topology in terms of a finite series of Renyi entropies of successively higher orders for the rows and columns. This sequence not only defines the topology uniquely, it does so as a sequence of successively smaller terms continuing topological information much like the Fourier expansion of sound waves of musical instruments. Secondly we were able to construct a network that consisted of nodes each of which is a network itself. This allows one to position a given network of interest in relation to known networks and potentially can lead to a beginning for the classification of networks in terms of the location of the network of interest in this space and with that position relative to reference networks. Thirdly, we have been able to show that the eigenvectors of the Markov matrices generated by the Lie algebra monoid reveal the complex structure of clusters along with extensive data on the profile of that structure. Finally, we have developed a method of generating a network where entities (such as elements, corporations, or people) are the nodes and where the connection matrix is defined in terms of multiple (possible weighted) properties of those entities. When we then study the clustering in these networks, it is highly revealing of the underlying structures. This algorithm has very extensive applicability due to its generality.

REFERENCES

Johnson, Joseph E, 1985 *Markov-Type Lie Groups in GL(n,R)* Journal of Mathematical Physics 26 (2) 252-257

Johnson, Joseph E. 2005 *Networks, Markov Lie Monoids, and Generalized Entropy, Computer Networks Security*, Third International Workshop on

Mathematical Methods, Models, and Architectures for Computer Network Security, St. Petersburg, Russia, Proceedings, 129-135US

Johnson, Joseph E., 2006 *Markov Lie Monoid Entropies as Network Metrics* MIT ICCS Conference on Networks & Complex Systems

Campbell, John William, 2014, *Network Analysis and Cluster Detection Using Markov Theory* M.S. Thesis, University of South Carolina

Johnson, Joseph E. 2012 *Methods and Systems for Determining Entropy Metrics for Networks* US Patent 8271412.