Weighted networks: the issue of dichotomization

Rodolfo Baggio

Master in Economics and Tourism
Bocconi University
via Röntgen, 1 - 20136 Milan, Italy
tel.: +39 0258365437 - fax: +39 0258365439
and
National Research Tomsk Polytechnic University
30 Lenin Avenue, Tomsk 634050, Russian Federation

email: rodolfo.baggio@unibocconi.it

==========
International Journal of Tourism Sciences (2019), forthcoming

Abstract

The purpose of this methodological paper is to explore one issue in the analysis of complex networks when weights are used to value edges. Often these weights are removed by setting a threshold and considering the links existing only if their value is higher than this. This practice allows then using simpler metrics as provided by many software packages. However, many common network properties, which often lead to specific interpretations, can be highly sensitive to the assumptions and thresholds used. By discussing a case, we show the differences that arise when dichotomizing a weighted network. We conclude that while unweighted networks can provide insights into some structural properties, the operation can be unnecessary and even detrimental for studying many features and processes when valued relational data are available.

Keywords

network analysis, weighted networks, dichotomization.

1 Introduction

The view of tourism as a complex phenomenon which embraces complex socio-economic systems is today well accepted and used for studying several characteristics of the domain, mainly for what concerns the status and the evolution of the systems that belong to the domain (Baggio, 2008; Noh, 2009; Speakman & Díaz Garay, 2016; Zahra & Ryan, 2005).

Among the many methodological approaches for the study of the complexity of many natural and artificial systems, network analysis has proved to be especially suitable, mainly when socio-economic phenomena are involved. The methods for analysing a network have seen an incredible development in the last decade. The basic idea is that any system, from the simplest to the most complex, can be modelled as a network in which the different elements are connected by some kind of link. Network analysis has proven to be a powerful and soundly based way for the assessing the structure and the dynamic behaviour of such systems and has provided tools for better studying and simulating many processes unfolding on and in a network (Barabási, 2016; Cimini et al., 2018). Practically all disciplines have been impacted and tourism is no exception (Baggio, 2017; Casanueva et al., 2016; van der Zee & Vanneste, 2015). Apart from studies concerning the structural characteristics of a system, different types of networked assemblies have been used to explore other topics such as the propensity to favour creativity and innovation (Baggio, 2014) or the evaluation of a destination's performance (Stienmetz & Fesenmaier, 2013), or the resilience to possible climate changes (Luthe & Wyss, 2016).

In this work we deal with one specific issue: dichotomization. When a network is weighted (i.e. its links are assigned a value, see the next section for a discussion), one way to use existing techniques and tools for calculating the measurements that are needed or desired is to scale it down and assume that a link exists only if its value is higher than a certain threshold. This allows to obtain a simple binary graph (where a link has value = 1 if it exists, 0 if not) and use one of the many software applications that handle these objects. However, as it is easily understandable, in this way much information is lost. The major problem is that that there is no accurate way for defining the threshold, but its value is left to the sensibility and the experience of the researcher about the object under investigation. The resulting networks, however, can be greatly different in structure form the original (Thomas & Blitzstein, 2017), and the situation can be even worse if individual actors' centralities are concerned (Eisenkraft, 2017).

In other words, even using "rigorous" methodologies (provided they exist), there is a high risk of having warped perceptions of the importance of certain elements (local or global) in the network and substantial losses of efficacy may arise when examining dynamic processes such as information and knowledge sharing or opinion formation.

The rest of this paper is organized as follows. After a short summary of the basics of network analysis, weighted measurements are discussed. A worked example, then, shows the differences obtained using different thresholds for the dichotomization of a network.

2 Networks and weighted networks: the main metrics

A network is an abstraction used to model a system. The basic components are nodes (also called vertices) and links (edges) connecting them. Both components can be used as *simple* items or be assigned some value (weight) that renders some specific characteristic such as size, value, cost, length and so on; in this case we speak of *weighted networks*.
Most investigations on networks (and processes on these networks) have focused on binary ties (i.e. simply existing or not). For these a wide variety of metrics have been proposed that render the basic structural and dynamic properties at global, intermediate or individual level (Barabási, 2016; da Fontoura Costa et al., 2007). Among these, well known measures are:

- *degree*: the number of connections each node of the network has to other nodes;
- *density*: number of actual connections between a set of nodes compared to the number of links if nodes were fully connected;
- *assortativity*: the correlation between nodal degrees;
- *average path length*: the average distance (shortest path) between any two nodes and *diameter*, the maximal path length in the network; for reducing the effects of outliers an more robust metric, the effective diameter, can be calculated as the 90^{th} percentile of all distances;
- *clustering coefficient*: the degree of non-homogeneity in the density of links;
- *closeness*: the capability of a node to "reach" any other node in the network;
- *betweenness*: the measure of the extent to which a vertex lies on the paths between others, thus acting as a bridge (or a bottleneck) between different parts of the network;
- *efficiency*: a measurement of the ease with which information flows;
- *modularity*: the extent of division in denser sub-networks, also called communities.

When the quantities refer to a single element (node) and are subject to some normalization (typically on the network's size) the term centrality is commonly employed.

For many of these measurements interesting and used quantities are the means and the forms of their statistical distribution or other measures that render the distribution of the values. For the degrees, for example, apart from the their distribution, the Gini index provides an immediate assessment of the heterogeneity of their values (Hu & Wang, 2008).

Little work has considered so far the addition of weights to nodes (among the few examples is Wiedermann et al., 2013), while the practice of weighting the connections existing between nodes is

relatively common mainly when the characteristics of social and economic systems (like a tourism destination, for example) are at play. Nonetheless, the literature has provided also, at least for the most important metrics, a "weighted" version, that takes into account the values assigned to the links (Barthélemy et al., 2005; Newman, 2004; Opsahl et al., 2010).

The outcomes from an analysis can be greatly different when weights are taken into account. It is relatively easy to see how the valuation of links can alter many of these measurements with respect to the unweighted measure. A simple example is shown in figure 1.

Let us suppose that the weights shown on panel B represent costs. The node with the highest degree k in A is B (k=4), while in network B is E since the total weight of the links k_w = 2 (for B k_w=1.4). If then we consider the shortest path l between B and C we have l=1 for the direct connection B-C (binary network A) and l=0.4 for the path B-A-C (weighted network B).

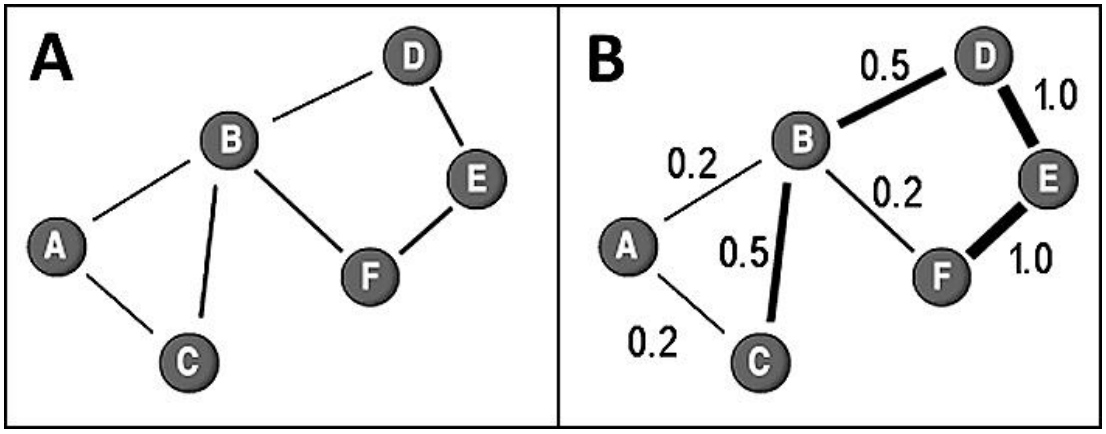

Figure 1 A binary network (A) and a weighted version of the same network (B)

In many cases the extension of a metric to a weighted version is trivial, in some cases it is more complicated or has not been "solved" yet (for example no good definition of *weighted density* still exists). The details of the formulas are skipped here for easing the readability of this paper, interested readers can find all details in the vast literature on the subject (Barabási, 2016; Barrat et al., 2004; Barthélemy et al., 2005; Newman, 2004, 2010; Opsahl et al., 2010).

Often, however, when examining a network with valued edges, researchers rely on some extemporary method to produce a binary replacement. The most common method is to dichotomize the values by choosing a threshold value, and assuming a link to exist if the threshold is exceeded. Examples of this practice exist in many domains, tourism included (for the latter a few recent examples are: Langle-Flores et al., 2017; Liu et al., 2017; Wäsche, 2015).

The fact that almost all the most used software packages available (such as UCINET or Pajek) do not handle, or handle only partially, weights is a possible cause for this diffused practice. For a correct treatment of all the aspects of a weighted network, however, one needs to resort to one of the commonly used software development environments (Matlab, Python, R etc.) that offer a wide range of libraries all including a wide variety of features.

3 An example: dichotomizing a weighted network.

In order to better understand the problems that can be found in dichotomizing a weighted network, let us discuss the following example. Data used are from the UK Department for Transport (https://roadtraffic.dft.gov.uk/) and refer to the Region of London. Data represent traffic figures (annual average daily flow in vehicles per day) for each junction to junction link on the major road network (the year considered is 2014). The main characteristics of the resulting network are shown in table 1. Here we report the metrics calculated for the weighted network and for its unweighted version. This comparison is frequently an initial step that allows a coarse but important assessment of the global topology of the network and the influence of the weights (Baggio et al., 2011; Estrada & Bodin, 2008).

Table 1. Basic characteristics of the network

	Weighted	**Unweighted**
Node count:	788	788
Link count:	1294	1294
Density:	---	0.004
Giant connected component:	100%	100%
Diameter:	521 043	16
Effective diameter (90%)	359 307.7	5.3
Ave path length:	107 499.7	3.95
Clustering coefficient:	0.015	0.123
Global efficiency:	0.000	0.287
Ave local efficiency:	0.000	0.143
Assortativity:	-0.117	-0.127
Modularity:	0.587	0.649
Gini index degrees:	0.649	0.518
Degree distribution exponent α:	3.31±0.32	2.2±0.05

The degree distributions, as for many other networks, follow a power law $N(k) \sim k^{-\alpha}$. This is considered to be a signature of the complexity of the system and to provide some hints on the possible formation mechanism of the network (Barabási, 2016; Cimini et al., 2018).

From the table it is possible to see how certain quantities are affected by the weighting scheme. In particular the heterogeneity (Gini coefficient and the exponent of the degree distributions), the clustering coefficient, the assortativity, and the efficiencies. All these are normalized quantities, so the comparison is easy.

As a second step we generate different binary versions of the network using three different thresholds and selecting: the top 75% weights (N75), the top 50% (N50) and the top 25% (N25). The networks are shown in figure 2 (for the full network ALL, the unweighted version is depicted). Table 2 shows the metrics for all networks (the full network the unweighted version).

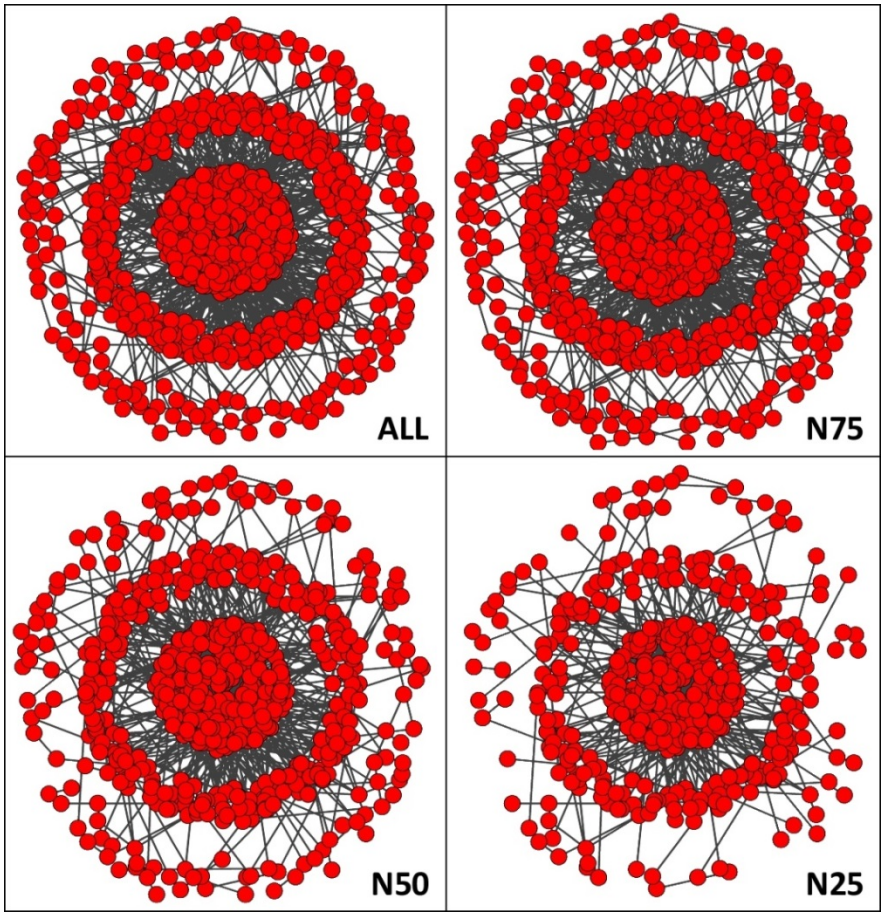

Figure 2. The four networks examined

Table 2. Metrics for the binary networks

	ALL	N75	N50	N25
Node count:	788	760	622	440
Link count:	1294	1235	879	548
Density:	0.004	0.004	0.005	0.006
Giant connected component:	100%	98%	92%	90%
No. of components	1	7	22	21
Diameter:	16	14	11	11

Effective diameter (90%)	5.3	5.2	5.2	4.9
Ave path length:	3.95	3.89	3.85	3.67
Clustering coefficient:	0.123	0.126	0.105	0.107
Global efficiency:	0.287	0.251	0.250	0.280
Ave local efficiency:	0.143	0.117	0.120	0.146
Assortativity:	-0.127	-0.130	-0.148	-0.181
Modularity:	0.649	0.647	0.644	0.670
No. of communities	21	17	27	21
Gini index degrees:	0.518	0.514	0.489	0.463
Degree distribution exponent:	2.20±0.05	4.10±0.63	3.71±0.48	4.50±0.52

Even a simple look at the picture shows quite dissimilar objects. However, it is only measuring the main quantities that we can really assess the differences. As it could have been expected, the dichotomized networks show a decreased heterogeneity (Gini index) in the degrees and a tendency to become fragmented (but as the degree distribution exponents show an increase "hubization"). This can affect quite severely the interpretation of the outcomes in a specific context. Apart from these global alterations, also the individual measurements tell different stories. Table 3 reports the rankings (top 10 nodes) of the networks' elements with respect to the most commonly used centralities. The node numeric ids have been kept constant across the different networks (i.e. a numeric label identifies the same node in all networks).

Table 2. Ranking (top 10 nodes) for centrality metrics in the different networks.

Rank	Degree					Clustering coefficient				
	WEI	ALL	N75	N50	N25	WEI	ALL	N75	N50	N25
1	18	18	18	18	18	5	5	5	5	5
2	1	23	23	23	1	403	31	31	31	31
3	23	86	1	1	23	225	56	56	34	34
4	148	1	86	85	184	766	79	79	42	42
5	85	155	155	86	86	457	97	97	56	66
6	6	96	96	155	50	133	100	100	58	74
7	184	50	98	21	85	171	105	105	79	79
8	30	93	184	98	21	337	133	133	100	114
9	86	98	50	184	98	531	141	141	124	133
10	50	184	93	93	6	377	146	146	133	134
Rank	**Closeness**					**Betweenness**				
	WEI	**ALL**	**N75**	**N50**	**N25**	**WEI**	**ALL**	**N75**	**N50**	**N25**
1	18	18	18	18	18	18	18	18	18	18
2	438	23	85	85	23	155	50	158	1	1
3	347	85	23	23	1	23	158	1	158	158
4	83	155	155	86	50	86	1	23	23	23
5	157	86	86	1	162	19	23	50	155	149

6	498	60	60	21	184	81	101	101	50	50
7	280	21	21	155	85	96	64	64	125	20
8	301	1	1	139	109	71	86	155	181	30
9	765	158	158	162	139	20	155	86	149	184
10	374	64	50	50	86	158	60	20	101	17

To summarize these measurements it is possible to use an "importance index" calculated as the geometric mean of all the basic (degree, clustering coefficient, betweenness and closeness) normalized metrics used (Sainaghi & Baggio, 2014). The rankings for the top 10 nodes are in table 3.

Table 3. Importance ranking for the networks

Rank	WEI	ALL	N75	N50	N25
1	18	18	18	18	18
2	30	23	23	23	1
3	23	86	155	86	23
4	65	155	96	17	50
5	50	21	86	21	86
6	17	96	1	144	17
7	1	1	21	98	389
8	6	98	98	85	21
9	86	50	85	1	30
10	148	123	60	155	85

The differences are evident. Mostly between the original weighted network (WEI) and the binary versions. A known and widely used measure for the agreement between different variables is the Kendall's coefficient of concordance *W*: a non-parametric statistic hat assesses agreement among ratings (*W* ranges from 0: no agreement to 1: complete agreement). In our case the test run over all the nodes that appear in all networks is $W = 0.408$ (with high significance: $p<<10^{-5}$), thus confirming quantitatively the visual impression. That is to say that there is an very high probability (almost a certainty) of misinterpreting the many of the nodes' positions in the network and mistake their "real" importance. This can be a big issue when analysing the behaviour and the properties of actors in a social or socio-economic network, such as those commonly studied in the tourism domain, and interpreting the outcomes.

Centrality metrics, moreover, do not only measure the "importance" of a node, but represent the expected values of certain kinds of node participation in network dynamic processes. As such they can have a deep influence on the unfolding of these processes (Barrat et al., 2008; Borgatti, 2005;

Pastor-Satorras et al., 2015). The contribution of a node to the global behaviour is not only shaped by the structure of the system but stems from the interplay between dynamics and structure, and the value of the centrality measures are an important parameter controlling the extent and the duration of many dynamic processes (Malliaros et al., 2016; Restrepo et al., 2006).

4 Concluding remarks

A network is a useful and important abstract model for understanding the structural and dynamic characteristics of a complex adaptive system such as those that belong to the tourism domain. In many cases these networks, besides making visible the connections between the different actors, possess more detailed characteristics usually rendered as values assigned to the links.

Their analysis is a bit more complicated and cannot always be conducted fully with standard software packages. In these cases researchers have often adopted a practice that consists of defining a threshold for the links' values and dichotomize the network considering links existing only if their value is higher than the threshold. This may result, as also shown by the example presented, in an alteration of many metrics and of the relative importance of their components.

Although a dichotomized unweighted (binary) network can provide interesting insights into some structural properties, this procedure can be unnecessary and even detrimental for studying many features and dynamic processes when valued relational data are available. This affects both the judgement on the global properties of the network and the centrality values for the nodes. Moreover, when dynamic processes are examined, the weights can greatly change the dynamics and result in quite different outcomes.

Obviously not all networks behave in the same way, and not all situations lead to the same outcomes as those presented in this paper. Different results might be obtained depending on the values of the weights and on their distribution across the links. However, the issue should induce a serious warning and push researchers to verify carefully the possible variations that could arise in the measurements.

Even if a dichotomization had a strong theoretical or methodological justification, the use of appropriate weighted techniques undoubtedly allow for more nuanced interpretations of the network characteristics.

References

Baggio, J. A., Salau, K., Janssen, M. A., Schoon, M. L., & Bodin, Ö. (2011). Landscape connectivity and predator–prey population dynamics. *Landscape Ecology, 26*(1), 33-45.

Baggio, R. (2008). Symptoms of complexity in a tourism system. *Tourism Analysis, 13*(1), 1-20.

Baggio, R. (2014). Creativity and the structure of tourism destination networks. *International Journal of Tourism Sciences, 14*(1), 137-154.

Baggio, R. (2017). Network science and tourism – the state of the art. *Tourism Review, 72*(1), 120-131.

Barabási, A. L. (2016). *Network science*. Cambridge, UK: Cambridge University Press.

Barrat, A., Barthélémy, M., & Vespignani, A. (2008). *Dynamical Processes on Complex Networks*. Cambridge: Cambridge University Press.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of the Sciences of the USA, 101*, 3747-3752.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2005). Characterization and modeling of weighted networks. *Physica A, 346* 34-43.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks, 27*(1), 55-71.

Casanueva, C., Gallego, Á., & García-Sánchez, M. R. (2016). Social network analysis in tourism. *Current Issues in Tourism, 19*(12), 1190–1209.

Cimini, G., Squartini, T., Saracco, F., Garlaschelli, D., Gabrielli, A., & Caldarelli, G. (2018). *The Statistical Physics of Real-World Networks* (arXiv:physics/1810.05095). Retrieved October 2018, from https://arxiv.org/abs/1810.05095.

da Fontoura Costa, L., Rodrigues, A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167-242.

Eisenkraft, N. (2017). Dichotomizing Network Data Can Change The Meaning of Actor Centrality. In G. Atinc (Ed.), *Academy of Management Proceedings* (Vol. 2016 (1), pp. no. 13383). Briarcliff Manor, NY: Academy of Management.

Estrada, E., & Bodin, Ö. (2008). Using network centrality measures to manage landscape connectivity. *Ecological Applications, 18*(7), 1810-1825.

Hu, H. B., & Wang, X. F. (2008). Unified index to quantifying heterogeneity of complex networks. *Physica A, 387*(14), 3769-3780.

Langle-Flores, A., Ocelík, P., & Pérez-Maqueo, O. (2017). The role of social networks in the sustainability transformation of Cabo Pulmo: A multiplex perspective. *Journal of Coastal Research, 77*(sp1), 134-142.

Liu, B., Huang, S. S., & Fu, H. (2017). An application of network analysis on tourist attractions: The case of Xinjiang, China. *Tourism Management, 58*, 132-141.

Luthe, T., & Wyss, R. (2016). Resilience to climate change in a cross-scale tourism governance context: a combined quantitative-qualitative network analysis. *Ecology and Society*, *21*(1).

Malliaros, F. D., Rossi, M. E. G., & Vazirgiannis, M. (2016). Locating influential nodes in complex networks. *Scientific reports, 6*, art. 19307.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E, 70*, 056131

Newman, M. E. J. (2010). *Networks - An introduction*. Oxford: Oxford University Press.

Noh, H. N. (2009). Enclave System and The Transition to Clans Tourism: Theoretical Perspectives Based on Complexity Science in Central and South America. *International Journal of Tourism Sciences, 9*(1), 1-22.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*(3), 245-251.

Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2015). Epidemic processes in complex networks. *Reviews of Modern Physics, 87*(3), 925-979.

Restrepo, J. G., Ott, E., & Hunt, B. R. (2006). Characterizing the dynamical importance of network nodes and links. *Physical Review Letters, 97*, art. 094102.

Sainaghi, R., & Baggio, R. (2014). Structural social capital and hotel performance: is there a link? *International Journal of Hospitality Management, 37*, 99-110.

Speakman, M., & Díaz Garay, A. (2016). Perspectives on tourism development planning in Acapulco: conventional methods and complexity theory. *International Journal of Tourism Sciences, 16*(4), 203-221.

Stienmetz, J. L., & Fesenmaier, D. R. (2013). Traveling the network: A proposal for destination performance metrics. *International Journal of Tourism Sciences, 13*(2), 57-75.

Thomas, A. C., & Blitzstein, J. K. (2017). *Valued ties tell fewer lies: Why not to dichotomize network edges with thresholds; 2011. Preprint. Available: . Accessed 27 June 2016.* (arXiv:1101.0788v2). Retrieved January, 2017, from https://arxiv.org/abs/1101.0788.

van der Zee, E., & Vanneste, D. (2015). Tourism networks unravelled; a review of the literature on networks in tourism management studies. *Tourism Management Perspectives, 15*, 46-56.

Wäsche, H. (2015). Interorganizational cooperation in sport tourism: A social network analysis. *Sport Management Review, 18*(4), 542-554.

Wiedermann, M., Donges, J. F., Heitzig, J., & Kurths, J. (2013). Node-weighted interacting network measures improve the representation of real-world complex systems. *Europhysics Letters, 102*(2), art. 28007.

Zahra, A., & Ryan, C. (2005). Complexity in Tourism Structures–the Embedded System of New Zealand's Regional Tourism Organisation. *International Journal of Tourism Sciences, 5*(1), 1-17.