A sample (chapter) from:

COMPLEX CONTAGION IN SOCIAL NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Christopher J. Cameron

May 2016

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 3

THE RELATIVE SHORTNESS OF LONG RANGE TIES

The preceding chapter identifies the emergence of clusters as a mechanism through which long range ties impact the adoption rates. Clusters increase contagion rates because they expose new nodes that are distance from the other regions of infection. Since clusters form from particular configurations of long range ties, the distribution of long range ties should impact the appearance of clusters. In particular, if long range ties are more likely between nodes that are close to each other, the network distance between the infected nodes and the nodes susceptible via long range ties would be tend to be smaller than on randomly permuted graphs. With this distribution of tie range, clusters would be more likely to appear near the periphery of the contagion perimeter, close to nodes already susceptible via short range ties. Crowded around the periphery of the main infected cluster, these new clusters would quickly merge with the seed cluster. Most new nodes exposed by the cluster would have been exposed to the seed cluster after a few waves of local spread. Since clusters reduce the time to saturation by increasing the number of nodes exposed to the contagion, these peripheral clusters would not be expected to reduce the saturation times. In this chapter, I consider the distribution of tie range in several empirical online social networks, compare the empirical distributions to those generated by the permutation algorithm used in prior complex contagion studies, and develop a permutation strategy that produces a more realistic tie range distribution while preserving other features of the Maslov—Sneppen permutation algorithm.

3.1 Previous findings

Though long range ties are an important element of sociological explanations (Grannovetter, 1973; Granovetter, 1983; Centola & Macy, 2007; Barash et al., 2012), the distribution of tie range has not been examined empirically. The computation of tie range has attracted significant attention of the past twenty years as the *second shortest path* problem in computer science but the literature focuses on the algorithm and not on empirical applications or surveys of social networks (Eppstein, 1994; Papaefthymiou, 1997; Li, Sun, & Chen, 2006; Kao, Chang, Wang, & Juan, 2011; Zhang & Nagamochi, 2012; Wu, 2013). A forthcoming work, Park (2016), aims to examine both the empirical distribution of tie range and possible mechanisms to produce the observed tie ranges in social networks. The present work is independent of Park (2016).

Most prior work examining the relationship between perturbed networks and contagion has relied on rewiring ties uniformly at random or creating new randomly directed ties (Watts & Strogatz, 1998; Centola et al., 2007; Centola & Macy, 2007; Barash et al., 2012). Kleinberg (2000) considered the distribution of long range ties in the context of decentralized search, introducing an alternative small world model. Recently, Ghasemiesfeh et al. (2013) considered contagion speed for stylized spatial networks with different distributions of weak ties. They consider a model where weak ties form at random – the Newman Watts model – and two models where the probability of weak tie formation depends on the proximity of the endpoints – the Kleinberg small world and hierarchical network models. Compared to contagions on networks with randomly formed long range ties, contagions on Kleinberg small world and hierarchical networks spread more slowly.

3.2 Tie range distribution in empirical social networks

To understand the empirical distribution of tie range, I selected an assortment of friendship and communication graphs drawn from online social networks, including networks from Flickr, LiveJournal, YouTube and Orkut from Mislove et al. (2007), the one hundred American college Facebook networks from Traud, Mucha, and Porter (2012) and Twitter data collected by the Social Dynamics Lab at Cornell University. These graphs range in size from thousands to millions of nodes and represent a variety of social networks organized around a variety of online activities. The computation of tie range is not a standard function of network analysis software so I implemented a custom tie range function using the iGraph C library (Csardi & Nepusz, 2006).

The proportion of long range ties in the surveyed networks ranges from .002 to .384. In general, larger networks have a larger proportion of long range ties but there is significant variation among networks of similar size. Table 3.1 and Table 3.2 report both the overall proportion of long range ties — P(LRT) — and the estimate *k* for the geometric distribution parameter for distribution of ties with range ≥ 3, which corresponds to the ratio between the number of ties at range *x* and the number of ties at range greater than *x*. For the purposes of this paper, it is sufficient to note that for the surveyed networks, most ties will be short range ties, most of long range ties will be length 3 and the frequency of ties with a particular range decreases dramatically as the range increases.

Table 3.1: Tie range statistics for large online social networks.

	Name	Nodes	Size	P(LRT)	k
Twitter:	Feb 2011	17,905	707,781	0.011	0.974
Egypt	Dec 2011	54,572	1,907,105	0.023	0.990
(all ties)	Jun 2012	94,140	3,029,650	0.038	0.988
	Aug 2013	109,082	3,425,747	0.081	0.981
Twitter:	Feb 2011	9,166	100,662	0.075	0.856
Egypt	Dec 2011	17,305	170,782	0.097	0.847
(mutual ties)	Jun 2012	28,805	211,671	0.144	0.758
	Aug 2013	41,879	235,483	0.205	0.629
Online	Orkut	2,997,354	113,910,527	0.134	0.914
Social	Flickr	1,192,171	13,779,153	0.054	0.774
Networks	LiveJournal	4,150,633	45,414,720	0.141	0.711
	YouTube	509,331	2,333,627	0.384	0.682

Table 3.2: Tie range summary for 100 Facebook networks.

Statistic	Nodes	Size	P(LRT)	k
Mean	12,070	469,838	0.017	0.967
St. Dev.	(9,067)	(363,540)	(0.011)	(0.016)
Min	762	16,651	0.002	0.920
Max	41,536	1,590,651	0.048	0.996

Overall, the surveyed networks have strikingly similar distributions of tie ranges — the pattern is qualitatively similar to a geometric distribution. Figures 3.1, 3.2, 3.3 show the proportion of ties with each range to illustrate the observed distribution. By definition, two is the minimum tie range for a graph without parallel edges. The ordinate axes is the logged proportion, which highlights the exponential decay in the number of ties at each range — the straighter the line, the more similar the distribution is to a geometric or negative binomial distribution. The empirical distributions are not a match for a geometric or negative binomial according to a goodness of fit test (Meyer, Zeileis, & Hornik, 2015); the number of ties with larger ranges tends to be larger than the expected counts. Future work might explore the interaction of tie range and network diameter, which constrains the range distribution. In any case, the discrepancies from the geometric distribution are in the higher ranges and represent only a small fraction of the long range ties.

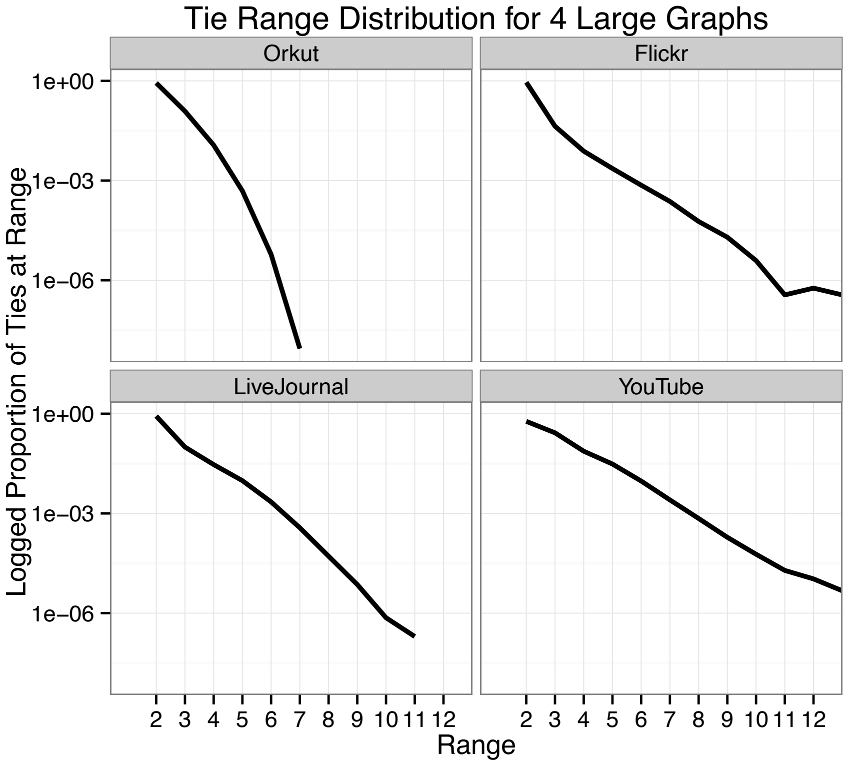

Figure 3.1: Percent of ties at each value of range for four large online social network graphs.

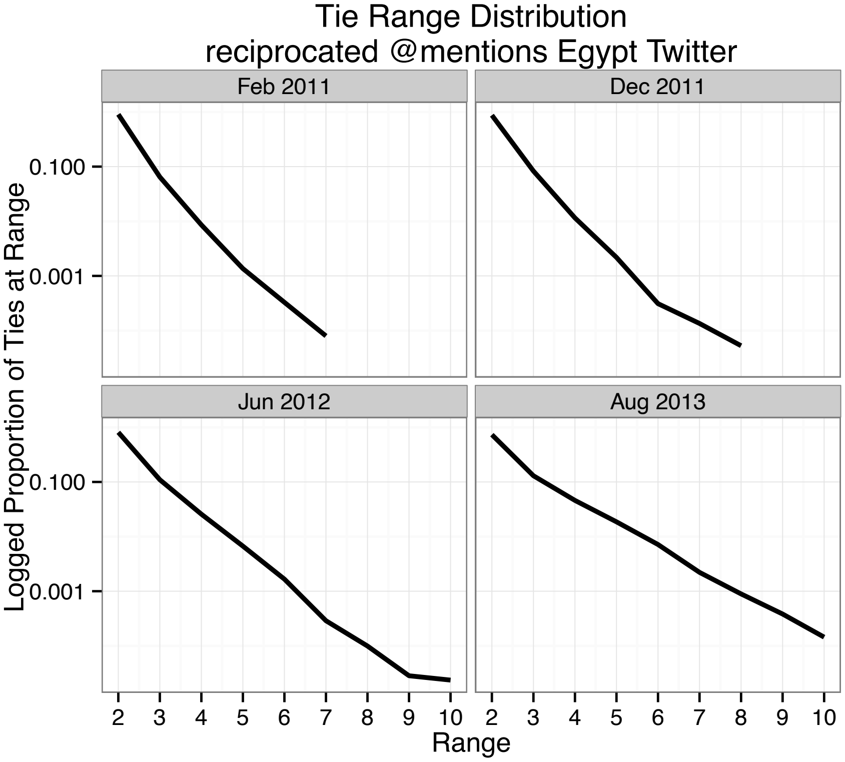

Figure 3.2: Percent of ties at each value of range in the reciprocated @mention network for Egyptian users active in four time periods.

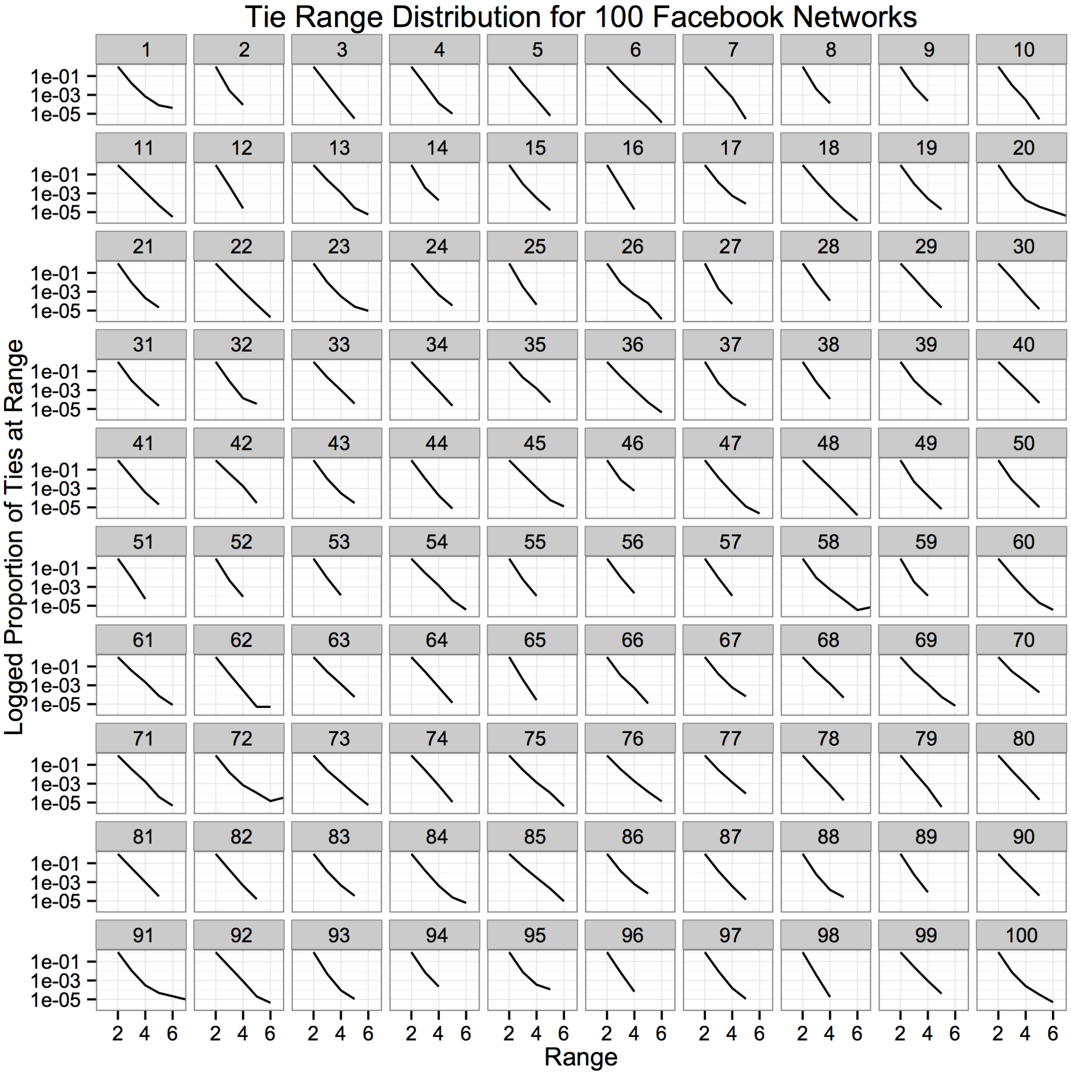

Figure 3.3: Percent of ties at each value of range in 100 Facebook friendship networks.

3.3 Tie range on MS perturbed lattice networks

Next, I examine the tie distribution that results from the perturbation algorithm introduced by Maslov and Sneppen (2002), sometimes called the double edge swap method. This approach has the benefit of preserving the degree of each node, thus preserving both the degree distribution and the density of the graph. Contagion processes are sensitive to both the shape of the degree distribution and the overall graph density, so Maslov—Sneppen (MS) rewiring can be used to examine the impact of tie perturbation while keeping other factors constant (Centola et al., 2007; Centola & Macy, 2007; Barash et al., 2012).

Figure 3.4 shows typical tie range distributions produced by MS rewiring at different levels of rewire. In the initial unperturbed lattice, all ties are range two. As these ties are randomly re-allocated to long range ties, the distribution of long range ties changes from flat to unimodal with a peak near the expected average path length, $\frac{\log N}{\log k}$ where N is number of nodes and k is the mean degree.

The convergence is apparent even at low levels of rewiring. This unimodal distribution is quite different from that observed in the empirical networks so it may be that the role of long range ties in the spread of contagions is different than generally recognized in the literature. In particular, if long range ties are not usually far-reaching, then the chance of spawning independent, locally-isolated clusters of infections becomes quite small. Instead, the contribution of long ranges ties will be focused near the boundary between infected and uninfected parts of the graph.

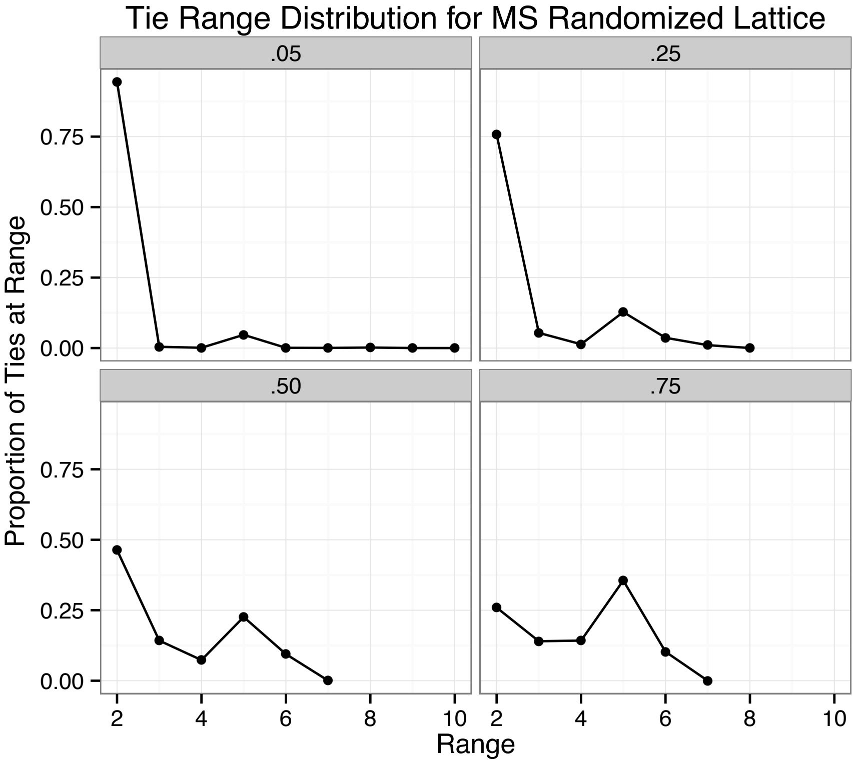

Figure 3.4: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired according to the Maslov–Sneppen method. Note that the axes are not logged.

3.4 Geometric biased double-edge-swap

A permutation strategy that is biased towards producing shorter long range ties would better match the empirical tie range distributions and could provide insight into how the complex contagion processes might unfold on more realistic networks.

Maslov—Sneppen rewiring (Maslov & Sneppen, 2002) operates by selecting two edges $A \rightarrow B$ and $C \rightarrow D$ uniformly at random and proposing two new edges by swapping the endpoints of the selected edges: $A \rightarrow D$ and $C \rightarrow B$. If either of the proposed edges already exists, then a new pair is selected at random. If the proposed edges do not exist, they are added to the graph and the originally selected edges are removed. In this way, they network is rewired while preserving the degree of each node.

I created more realistic tie range distributions by modifying the way edges are selected for rewiring. Note that since all the lattice networks are degree regular and undirected, choosing a node at random and then a random edge incident to that node is the same as selecting an edge at random. The selection proceeds as follows:

1. Choose a source node at random.
2. Choose a distance *d* by taking a random draw from a geometric distribution with parameter = 0.95.
3. Choose a target node at random from the set of nodes with the shortest path from source node equal to *d* if such a node exists, otherwise restart.
4. Choose a one random edge incident to the source node
5. Choose a second random edge incident to the target node.
6. Apply the Maslov—Sneppen swap method to the two selected edges.

While not particularly efficient, this iterative approach generates tie range distributions with the same qualitative properties as the empirical distributions, particularly for the proportions of long range ties observed in the empirical networks. Figure 3.6 uses the same axes as the empirical networks and the downward trend is apparent up until the proportion of long range ties exceeds 0.50. This geometric biased rewire procedure (GEO) also creates distributions quite different from the the original Maslov—Sneppen permutation. Figure 3.5 uses the same axes as Figure 3.4 to facilitation the comparison of the two permutation approaches. Most notably the geometric biased avoids the unimodal shape until there are a large fraction of long range ties and the maintains the scarcity of far reaching long range ties in relative to the shorter range ties.

The GEO permutation strategy produces graphs that are very similar to MS permutation in many respects. Both permutations preserve the number of edges and the degree distribution. Importantly, the edge overlap distribution is nearly the same in both graphs. Overlap is the number of neighbors the endpoints of an edge have in common – in other words, it is the width of the bridge between neighborhoods. Centola and Macy (2007) showed the existence of wide bridges was necessary for complex contagions to spread, so it is important that the permutation techniques manipulate overlap to a similar degree for the results to be comparable. Figure 3.7 shows the distribution of overlap for random (MS) and GEO rewired networks. The number of wide bridges between neighborhoods is very similar though slightly lower on GEO permuted lattices. Overall, any differences in the outcome of contagion processes are more attributable to large differences in tie range distribution rather than small differences in overlap.

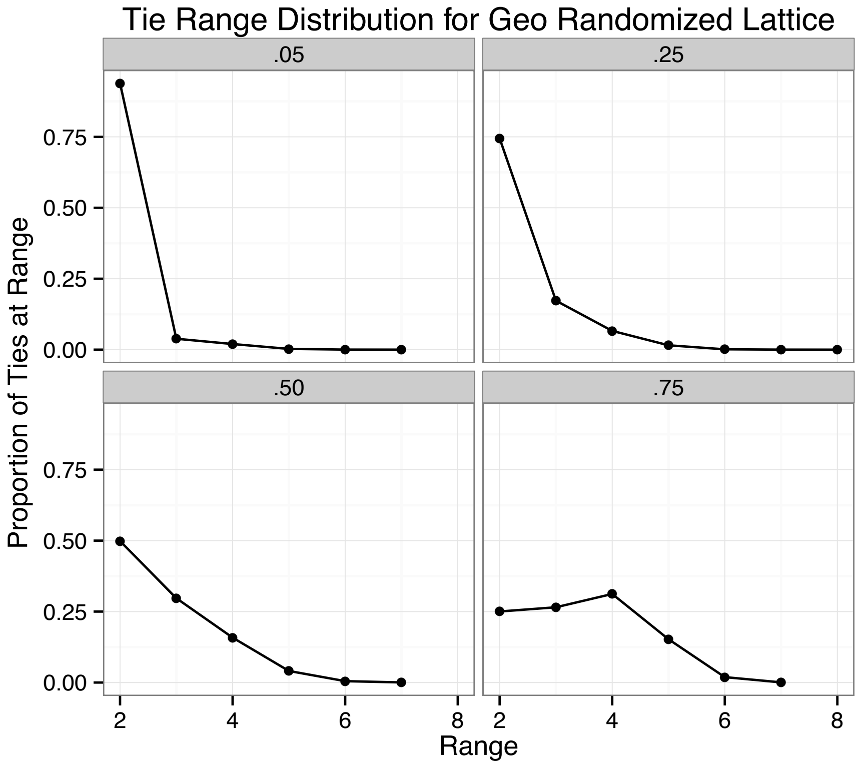

Figure 3.5: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired to favor the creation of shorter long range ties. Note that the axes are not logged.

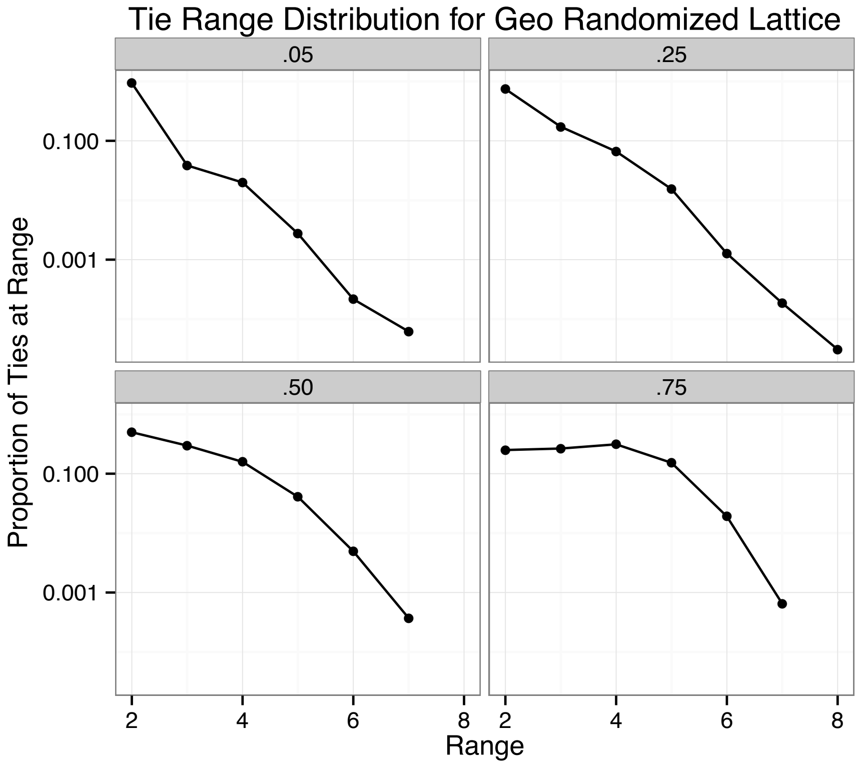

Figure 3.6: Percent of ties at each value of range for graphs with the proportion of long range ties equal to .05, .25, .50 and .75. These lattice graphs were rewired to favor the creation of shorter long range ties. The ordinate axis is logged to facilitate comparison with the empirical networks.

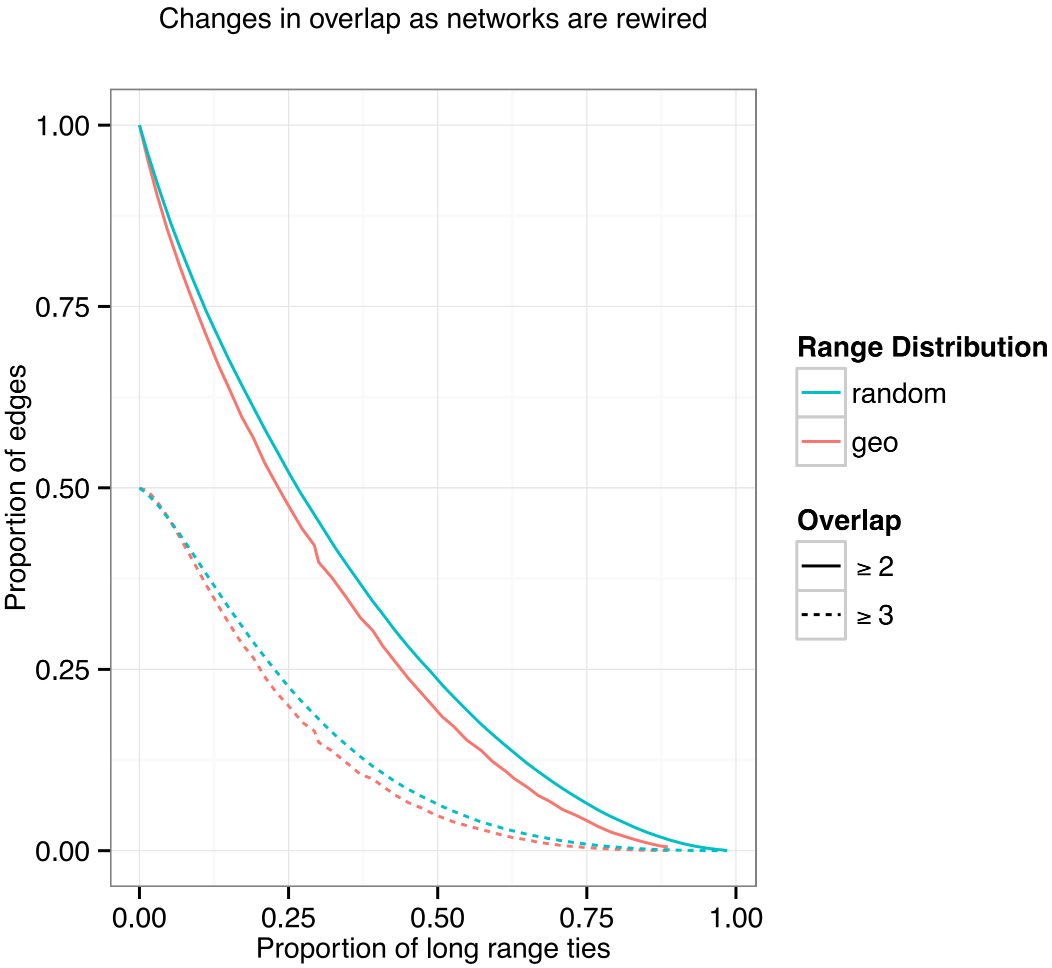

Figure 3.7: Comparison of neighborhood overlap under MS and GEO permuted 8-regular Moore lattice.

3.5 Discussion

A survey of several online social networks revealed that empirical tie range distributions are qualitatively similar in that most ties are range two and the preponderance of long range ties are range 3. The MS permutation used in previous simulations of complex contagion produces tie range distributions with fewer range three ties and more ties at the higher values of range than is typical of empirical networks. The analysis from Chapter 2 and a few examples from the literature show that the distribution of tie range can impact the rate a contagion spreads through the network so the discrepancy between MS and empirical networks could produce misleading model results. To support further investigation, I proposed the GEO permutation which produces a more realistic distribution of tie ranges while retaining the desirable properties of MS permutation.