Eigenvector-Centrality Dominance and Diffusion Processes

Brian Xu[1] and Hamed Hekmat[2]

[1]Department of Mathematics, Stanford University
[2]Department of Computer Science, Stanford University

June 9, 2025

Abstract

Inspired by empirical observations regarding microfinance diffusion in villages in India from Banerjee et al. [2013] that highlight eigenvector centrality as a better predictor of influence than degree centrality, we propose two novel diffusion models, LAB (Limited Attention and Branching) as well as LABC (LAB with Convincingness). LAB builds upon the existing Independent Cascade (IC) model by introducing a bounded attention mechanism to reflect attention scarcity and cognitive constraints, limiting how many neighbors an activated node attempts to influence, while LABC further extend this by introducing a convincingness parameter as a function of a given node's characteristics within the graph. We explore theoretical implications of such models on toy graphs while also conducting extensive experiments using Monte Carlo simulations to compare these models of diffusion to the classical IC and Linear Threshold (LT) models, examining whether certain conditions are more likely to bias eigenvector centrality (and lead to eigenvector centrality dominance, or ECD). Overall, we discuss how crucial graph structure and the design/selection of one's diffusion model rule can be for the optimal seeding/influence maximization problem, in tandem with heuristics such as eigenvector centrality or degree centrality.

1 Introduction

In network/graph theory, several models of diffusion have been devised to capture and reflect empirical properties of real-world diffusion processes. Such models are often applied to the optimal seeding problem, posing the question of how to effectively predict which seeds have potential to yield high diffusion spread, relevant for many applications such as viral marketing, social campaigning, information diffusion, and epidemiology. In pursuit of which node characteristics can serve as effective heuristics for seed quality, theorists have often used various measures of centrality such as degree centrality or eigenvector centrality.

It remains to be entirely clear in what conditions/situations some measures of centrality perform better than others. For example, Banerjee et al. [2013] examined the relationship between various centrality measures and total diffusion within the context of microfinance loan program participation in villages in India. They posit the idea of communication centrality or diffusion centrality, closely related to the eigenvector centrality of a given node, as an effective predictor of seed quality/diffusion spread within this particular empirical setting.

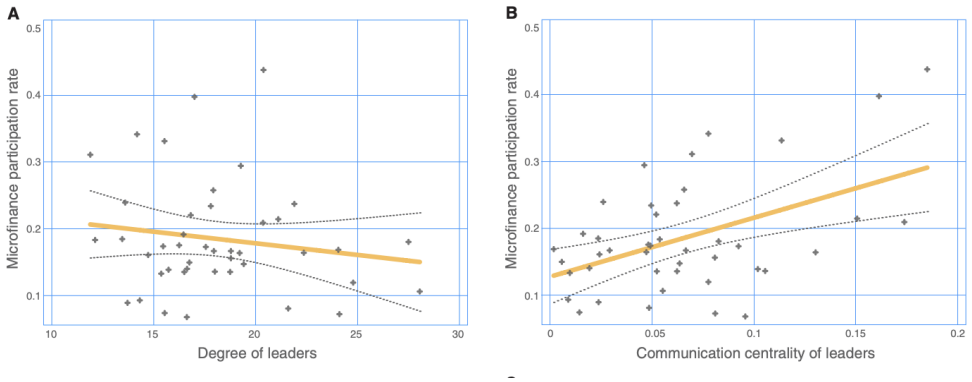


Figure 1: Comparing correlation of centrality measures with microfinance participation rate in Banerjee et al. [2013]

However, such empirical findings can be heavily dependent on the design of the diffusion model of choice, and many leading theoretical models of diffusion contain assumptions that stray from real-world processes. Independent cascade (IC) and linear threshold (LT) models in particular are two prominent choices in the field that assume that each node interacts with every single one of its neighbors, something that often isn't true when examining real-world dynamics and interactions.

In hopes of further exploring the underlying conditions that contribute to seed quality, we examine two extensions to the IC model, which we call Limited Attention and Branching (LAB) and Limited Attention and Branching with Convincingness (LABC). These extensions are intended to reflect the fact that in empirical networks, most nodes won't have the capacity to consistently and meaningfully engage with all of their connections. Additionally, we begin to investigate how graph structure can alter which heuristics most effectively predict diffusion spread/seed quality. We do so by first proposing some theoretical observations on toy graphs. Lastly, we discuss empirical findings from simulated diffusion processes on real-world graph datasets of varying sizes and network structures.

2 Related Work

Both the IC and LT models of diffusion were put forth by Kempe et al. [2003]. As a brief summary, the IC model involves every active node in a given round attempting to activate each of its neighbors independently, oftentimes with a homogenous probability of success. The LT model, on the other hand, is characterized by each node activating if enough of their neighbors (characterized by a weighted ratio) have already been activated. Additionally accomplished in this seminal work, the authors began to formalize the influence maximization problem, crucially characterizing the application with either of these two models of diffusion as a case of submodular function maximization which allowed them to demonstrate that a greedy algorithm yields a good approximate solution.

Much progress has since been made in exploring the optimal seeding/influence maximization problem. While a lot of this work involves demonstrating improvements to earlier approximate algorithms, many theorists are focusing on determining effective heuristics. Chen et al. [2010], for example, explored a heuristic that served as a significantly improved predictor of large spread. Most recently, Akbarpour et al. [2021] explores the problem from a different angle, arguing that significant boosts in diffusion can be accomplished from simply adding a few seeds in comparison to finding the optimal seeds.

The aforementioned Banerjee et al. [2013] played a significantly role in demonstrating how empirical findings may differ from the behavior of diffusion processes as predicted by theoretical models. By raising many questions about whether such observations pertained to unique characteristics of the dynamics of the villages in the dataset (such as graph structure or information dissemination behaviors), it sparked many of the original questions that inspired this project. Another relevant work that significantly contributed to our initial idea is the gossip model from Banerjee et al. [2019]. By hypothesizing that individuals that are considered to be "good gossips" by their community may serve as effective seeds, their empirical findings demonstrated that certain nodes placed at key positions in a graph (depending on graph structure) often yield significantly improved diffusion spread when compared to the nodes that have the most connections, emphasizing both the idea that quality of connections may matter more than quantity as well as the notion that graph structure plays a major role.

Researchers in many fields have also studied the intuitive notion that whether due to limited attention spans or some other resource constraint, we're unlikely to interact with everyone we're connected to. Within the realm of behavioral psychology and cognitive science for example, Huberman et al. [2008] points out that attention scarcity and the chaos of daily life limit Twitter users' interactions with one another, leading to a sparse underlying network (much smaller than the actual graph of followers) responsible for the vast majority of engagement and contributions. Weng et al. [2012] examined how information saturation (in terms of meme quantity and variety) restricts the amount of content that can truly go viral, arguing this is primarily caused by limited attention, and Aral and Van Alstyne [2011] crucially argued that an individual can only actively maintain so many strong ties at once due to cognitive constraints. In applying such ideas to modify existing diffusion models, some works restricted processes to ensure nodes don't interact with every single one of their neighbors, such as Boyd et al. [2006] which introduced a gossip constraint where each activated node only attempts to spread a rumor to one randomly selected neighbor. We develop LAB and LABC, our extensions of existing models of diffusion, with these considerations in mind.

3 Definitions

Our model follows in the tradition of the seeding literature. We suppose that there is an undirected graph $G = (V, E)$ in which we care about seeding a subset of nodes $I \subseteq V$. The diffusion then originates from these nodes, spreading through edges in the graph according to some stochastic process. Since we care about the group of nodes that the process will eventually diffuse to, which we also call *infected* nodes, it is useful to think of diffusion σ as mapping from the set of seeded nodes to a set (possibly random) of eventually infected nodes $\sigma(I) \subseteq V$.

Definition 3.1 (Diffusion Rules)**.** *A **diffusion rule** σ on G is a random variable $\sigma : \Omega_G \times 2^V \to 2^V$ mapping a set of source nodes to the set of all potentially infected nodes where $(\Omega_G, \mathcal{F}_G, \mathbb{P}_G)$ is a probability space that depends on the graph G.*

In our case, we care about seeding individual nodes and unearthing the correlations of this diffusion structures with eigenvector and degree centrality, so we can abbreviate

$$\sigma(\omega, \{i\}) = \omega(i) \text{ and with abuse of notation } \sigma(i) = |\sigma(i)|.$$

It will be clear when we refer to the set or to the magnitude of the set. First, we give a simple example to illustrate how well-known diffusion models fit into this framework.

Example 3.1 (One Round of Homogeneous Independent Cascade σ_{IC})**.** *When a node is seeded in the Homogeneous Independent Cascade diffusion rule, it contacts all of its neighbors in the first*

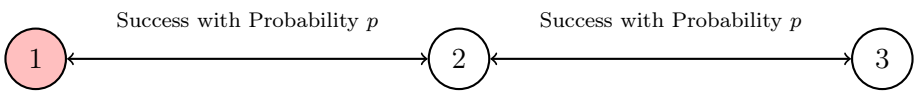


Figure 2: Example with Independent Cascade

round and flips a coin with probability p for each edge. Each heads results in a successful infection and spreads the process to the neighbor of the associated edge. Therefore, we may write the graph-dependent probability space here as the two coin-flip space given by

$$\Omega = \{0,1\}^2$$
$$\mathcal{F} = 2^{\Omega}$$
$$\mathbb{P}(b_1, b_2) = \prod_{i=1}^{2} p^{b_i}(1-p)^{1-b_i}.$$

Then, consider what happens when different nodes are seeded:

1. ***Seed Node 1:*** *Then, only the connection between 1 and 2 matters, so we have that*

$$\sigma_{IC}((b_1, b_2), \{1\}) = \begin{cases} \{1,2\} & b_1 = 1 \\ \{1\} & otherwise \end{cases}.$$

2. ***Seed Node 3:*** *Similarly, we get a symmetric argument to seeding node 1:*

$$\sigma_{IC}((b_1, b_2), \{3\}) = \begin{cases} \{2,3\} & b_2 = 1 \\ \{3\} & otherwise \end{cases}.$$

3. ***Seed Node 2:*** *Here, based on the values of b_1 and b_2, we can have any of 4 diffusion results happen:*

$$\sigma_{IC}((b_1, b_2), \{2\}) = \begin{cases} \{1,2,3\} & b_1 = b_2 = 1 \\ \{1,2\} & b_1 = 1, b_2 = 0 \\ \{2,3\} & b_1 = 0, b_2 = 1 \\ \{2\} & b_1 = b_2 = 0 \end{cases}.$$

We can summarize the IC diffusion rule originating from a single seed node as the following piecewise function:

$$\sigma_{IC}((b_1, b_2), \{i\}) = \begin{cases} \{1,2,3\} & i = 2, b_1 = b_2 = 1 \\ \{1,2\} & b_1 = 1 \text{ and } (i = 1 \text{ or } i = 2 \text{ and } b_2 = 0) \\ \{2,3\} & b_2 = 1 \text{ and } (i = 3 \text{ or } i = 2 \text{ and } b_1 = 0) \\ \{i\} & otherwise \end{cases}.$$

Before introducing LAB and LABC, we briefly describe the model logistics and probability space for IC and LT diffusion.

Definition 3.2 (Independent Cascade)**.** *In **Independent Cascade** $\sigma_{\mathbf{IC}}$, each newly informed node i (seeded or diffused to) successfully spreads information to its neighbor j independently with probability p_{ij}. The process ends when no new nodes become informed.*

Definition 3.3 (Linear Threshold)**.** *In* ***Linear Threshold*** $\sigma_{\mathbf{LT}}$*, each node i has a threshold v_i and weights over on edges b_{ij}. At every time t, a node i becomes informed if*

$$\sum_{j \in I_{t-1} \cap N_i} b_{ij} > v_i$$

The process ends when no new nodes become informed.

Now, we can introduce our models of diffusion.

Definition 3.4 (Limited Attention and Branching (LAB))**.** *In* ***LAB****, newly activated nodes attempt to spread information to a* ***limited amount of neighbors****, which can be deterministic or random, i.e. $N \sim \pi$ where π is an* ***offspring distribution*** *with bounded, positive support, i.e. $(N \geq 1) = 1$ and $(N \leq k) = 1$ for some $k \in \mathbb{N}$. Usually, we take $k \ll$ Average Degree. If the node has less than N connections, then it will try each of its N connections once. Each information spread attempt has a homogeneous probability of success $\beta \in (0, 1]$.*

Definition 3.5 (Limited Attention and Branching with Convincingness (LABC))**.** *As an extension to LAB,* ***LABC*** *operates functionally the same way, except each node has a convincingness probability to replace the homogenous β values from before. This parameter is computed as a function of eigenvector centrality to further emphasize that nodes with higher quality connections may be more effective in convincing their neighbors. We compute these values for each node by dividing its eigenvector centrality by the max eigenvector centrality from the whole graph:*

$$Conv(i) = \frac{c(i)}{max_{i \in V}(c(i))}$$

Lastly, we review the definitions of degree and eigenvector centrality, and we proceed to introduce a simple metric to assess whether a particular diffusion model and/or graph condition implies that eigenvector centrality is more correlated with diffusion spread than degree centrality is.

Definition 3.6 (Degree centrality)**.** ***Degree centrality*** *is given by*

$$d(i) = |\{(i, j) \in E\}|$$

Definition 3.7 (Eigenvector centrality)**.** ***Eigenvector centrality*** *is given by*

$$c = \frac{1}{\lambda_1} Ac$$

Definition 3.8. *A diffusion rule σ on G exhibits* ***eigenvector centrality dominance (ECD)*** *iff*

$$\text{Corr}\big([\sigma(i)], c(i)\big) \; > \; \text{Corr}\big([\sigma(i)], d(i)\big).$$

For a random graph model (although we won't be analyzing any), we say ECD holds w.h.p. *if the inequality is satisfied with probability $1 - o(1)$ as $n \to \infty$.*

4 Theoretical Insights

Last section, we defined what we mean by Eigenvector Centrality Dominance (ECD), a concept that is closely related to the findings of Banerjee et al. [2013]. In this section, we'll see that ECD is a joint property of the decision of diffusion rule as well as the graph G. We will show this with

propositions that show that 1. some graphs cannot evoke ECD from any diffusion rule and 2. some graphs can evoke ECD given certain conditions are true. What we learn from these propositions is that ECD is a very sensitive property of a network diffusion system—only when "things go right" can ECD come out as an empirical phenomenon.

The first proposition we present shows that sometimes ECD is impossible. This is accomplished by showing that eigenvector centrality and degree centrality can sometimes align, implying that ECD requires centrality and degree to be meaningfully different concepts.

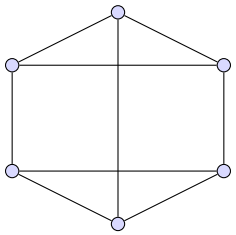


Figure 3: 3-regular graph with 6 nodes

Proposition 4.1 (Regular Graphs Never Exhibit ECD)**.** *If G is a k-regular graph, no diffusion rule σ can generate ECD.*

Proof. If G is a k-regular graph, we have that $d(i) = d(j) = k$ for all nodes $i, j \in V$. Similarly, the Perron eigenvalue is k and all the eigenvector centralities are equivalent. Thus, we have that $c(i) = c(j) = 1$. We can then see that $k \cdot c(i) = d(i)$ so for any diffusion process σ,

$$\text{Corr}\big(\mathbb{E}[\sigma(i)], c(i)\big) \;=\; \text{Corr}\big(\mathbb{E}[\sigma(i)], d(i)\big).$$

Thus, k-regular graphs never exhibit ECD. □

While we learn theoretically that ECD may not ever be evokable, the empirical reality of this proposition is that eigenvector centrality based targeting is then equivalent to degree centrality based targeting, so we don't learn much about when to use either heuristic in practice.

Next, we show that star graphs can exhibit ECD under specific conditions. In particular, μ, the mean of the offspring distribution, needs to be small enough so that LAB is not overly biasing large degree nodes (the hub).

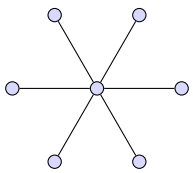


Figure 4: Star Graph S_7 with 6 peripheral nodes and one hub node

Proposition 4.2 (Star Graphs Can Exhibit ECD)**.** *Let G be a star graph S_{m+1}. Furthermore, suppose our diffusion process is σ_{LAB} with μ the mean of the offspring distribution and β the probability of infection. If*

$$\mu < \frac{1}{1 - \beta \cdot \left[\frac{m-1}{m}\right]}$$

then the LAB process on G exhibits ECD.

Proof. We can directly compute the expected diffusion sizes for the two types of nodes:

- Periphery nodes: With β chance the diffusion makes it into the hub node. Then, from the hub node, the chance that the process diffuses to an uninfected node is $\beta \cdot \frac{\mu}{m}$. Thus, the expected number of additional peripheral infections is $(m-1) \cdot \left(\beta \cdot \frac{\mu}{m}\right)$. So,

$$\mathbb{E}[\sigma_{LAB}(\text{Periphery})] = (1-\beta) + \beta \cdot \left(\beta\mu \left[\frac{m-1}{m}\right] + 2\right) = 1 + \beta + \beta^2\mu \left[\frac{m-1}{m}\right].$$

- Hub node: From before, we have that the chance a hub node infects a periphery node is $\beta \cdot \frac{\mu}{m}$, so linearity of expectation gives

$$\mathbb{E}[\sigma_{LAB}(\text{Hub})] = m \cdot \left(\beta \cdot \frac{\mu}{m}\right) + 1 = \beta\mu + 1.$$

Note that

$$\mu < \frac{1}{1 - \beta \cdot \left[\frac{m-1}{m}\right]} \implies \beta\left(1 - \mu + \mu\beta \cdot \left[\frac{m-1}{m}\right]\right) > 0$$

Thus $\beta + \beta^2\mu \cdot \left[\frac{m-1}{m}\right] > \beta\mu$ and therefore $\mathbb{E}[\sigma_{LAB}(\text{Periphery})] > \mathbb{E}[\sigma_{LAB}(\text{Hub})]$. To conclude, note that

$$d(\text{Hub}) = m \cdot d(\text{Periphery})$$

whereas

$$c(\text{Hub}) = \sqrt{m} \cdot c(\text{Periphery}).$$

Thus, correlation is higher with eigenvector centrality because $\sqrt{m} < m$ whereas the ratio between the spread from the hub node and the spread from a periphery node is less than 1. Hence, we have ECD for Star Graphs. □

We can interpret the results of the proposition and gain some insights into ECD and LAB. First, we know that as μ increases, the LAB process approaches the IC process. Thus, to obtain an upper bound on μ implies that we cannot let μ go to infinity, AKA use the IC process, if we want to obtain ECD on the Star Graph. In other words, the limited attention part of the LAB process is highly critical for generating ECD.

Second, when β or m increases, we get that the upper bound on μ for ECD grows. As β grows, the gap between the IC process and LAB process grows, as captured by the expected number of nodes that are infected (if we keep μ fixed). Thus, the LAB process can afford to "infect" more nodes when the conditions also point to IC and LAB being extremely different.

From this section, we have uncovered how ECD is a joint property of both the graph and the diffusion process. In some cases, ECD is impossible, whereas in others LAB can bring out ECD where others cannot. To some extent, the limited attention dynamics of the process are critical in generating the results that empirically see. In the next section, we extend our intuitions here to various real world graphs in order to measure correaltions and discover trends among the prevalence of ECD in simulations.

5 Empirical Evaluation

In the following section, we describe the simulated experiments we conducted on real-world graph datasets from Stanford Network Analysis Project (SNAP) and present some preliminary results and observations.

5.1 Datasets

We leveraged SNAP's broad array of large network datasets, seeking to capture a variety of network structures. In the table below, we present some overarching properties of the graph datasets used in our simulations. These graphs capture a variety of network/interaction types. For example, facebook_combined and email-Eu-core cover two different forms of using the internet to engage with others. wiki-Vote is derived from Wikipedia internal voting procedures, capturing a very different dynamic. p2p-Gnutella08 is a dataset capturing peer-to-peer filesharing interactions, and lastly, ca-GrQc comes from academic citations within the general relativity literature.

Graph	Nodes	Edges	ACC	Num Triangles	Fraction Closed
facebook_combined	4039	88234	0.6055	1612010	0.2647
ca-GrQc	5242	14496	0.5296	48260	0.3619
wiki-Vote	7115	103689	0.1409	608389	0.04564
p2p-Gnutella08	6301	20777	0.0109	2383	0.006983
email-Eu-core	1005	25571	0.3994	105461	0.1085

The following table represents more thorough statistics regarding the distribution of node degrees (degree centrality values) within each dataset.

Graph	Min	Max	Mean	Median	STD	Gini Coeff
facebook_combined	1	1045	43.69	25	52.414	0.541
ca-GrQc	1	81	5.53	3	7.918	0.555
wiki-Vote	1	1065	28.32	4	57.574	0.748
p2p-Gnutella08	1	97	6.59	3	8.54	0.524
email-Eu-core	1	347	33.25	23	37.313	0.540

Lastly, the following Jaccard similarity values help capture the overlap between the top 20% of nodes in terms of degree centrality vs the top 20% of nodes in terms of eigenvector centrality in each of our datasets.

Graph	Jaccard Similarity
facebook_combined	0.312
ca-GrQc	0.328
wiki-Vote	0.831
p2p-Gnutella08	0.418
email-Eu-core	0.74

5.2 Experimental Setup

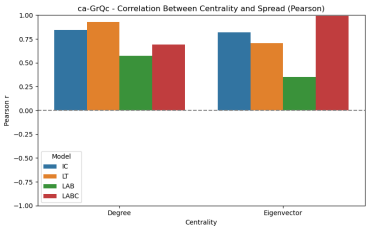
For each graph, we executed the following procedure to obtain simulation results.

```
Load graph from SNAP
Calculate network statistics

Select random seeds (roughly N/10)
Set parameters
    IC probability = 0.1
    LAB/LABC
```

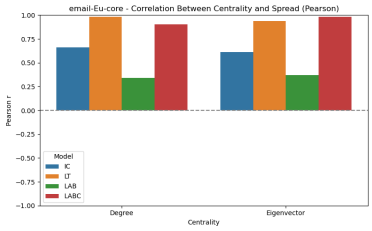
```
        mu = 2
        beta = 0.5
        k << avg(degree)

For each diffusion model (IC, LT, LAB, LABC)...
    For each randomly sampled initial seed node...
        Run Monte-Carlo simulation with 300 iterations
        Store average spread
    Compute descriptive statistics across seeds
        Spread (min, max, mean, median, std, gini coefficient)
        Correlations between degree centrality and spread
        Correlations between eigenvector centrality and spread
        Ratios for testing ECD
Generate plots
```

5.3 Results

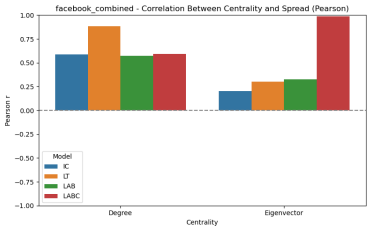
The initial bar charts below present correlations between degree/eigenvector centrality and diffusion spread across the various diffusion models. In any scenario where the right bar of a particular color is higher than the left, we know that model of diffusion yielded ECD on that specific dataset.



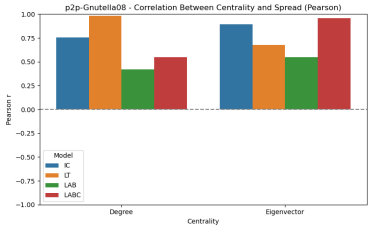
(a) ca-GrQc



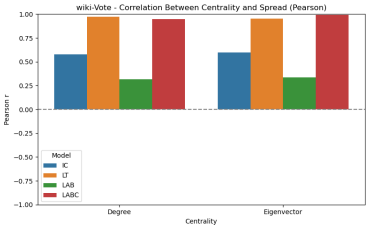
(b) email-Eu-core



(c) facebook_combined



(d) p2p-Gnutella08



(e) wiki-Vote

Figure 5: Correlation bar charts across all 4 models on 5 datasets

Next, we present a table that contains the correlation values from our experiments. For each dataset and diffusion model, we computed the correlation between degree centrality and diffusion spread as well as the correlation between eigenvector centrality and diffusion spread. We also included the ratio of the latter correlation over the former to quickly assess whether a particular set of circumstances yielded ECD.

Graph	Model	Corr(d vs spread)	Corr(c vs spread)	Ratio
facebook_combined	IC	0.588	0.205	0.348
	LT	0.882	0.299	0.339
	LAB	0.573	0.327	0.570
	LABC	0.590	0.989	1.676
ca-GrQc	IC	0.841	0.818	0.972
	LT	0.930	0.704	0.757
	LAB	0.574	0.350	0.609
	LABC	0.689	0.990	1.437
wiki-Vote	IC	0.576	0.599	1.039
	LT	0.0.973	0.0.952	0.978
	LAB	0.315	0.337	1.067
	LABC	0.949	0.991	1.045
p2p-Gnutella08	IC	0.755	0.893	1.183
	LT	0.981	0.676	0.689
	LAB	0.418	0.546	1.308
	LABC	0.546	0.958	1.753
email-Eu-core	IC	0.664	0.613	0.923
	LT	0.981	0.938	0.956
	LAB	0.339	0.370	1.092
	LABC	0.901	0.981	1.089

Figure 6: Correlation values and ratios for 4 models across 5 datasets

We've also included additional figures (scatterplots with lines of best fit) in the appendix for further reference.

6 Discussion

At a higher level, our results validate our initial intuitions. We hypothesized that whether a graph will exhibit ECD when tested using a particular diffusion process not only depends on the specifics of that diffusion process rule but also on the graph's structure in the first place. In this section, we provide some observations and reflections that stem from the empirical results of our simulations, as well as exploring some of the limitations of our work.

It's worth first highlighting some of the key ways in which our graph datasets differ from one another. For example, while all have between 1000 and 10000 nodes, facebook_combined, wiki-Vote, and email-Eu-core have significantly higher average node degrees than ca-GrQc or p2p-Gnutella08. This likely pertains to the kind of interactions being captured by the two sets, as the datasets with higher node degrees are derived from social platform interactions online (Facebook, Wikipedia, email), whereas the other two come from settings with a much lower volume of interactions (academic citations in general relativity, peer-to-peer file sharing). Additionally, we have a broad range in average clustering coefficient values, which likely implies that overall graph structures differ significantly. Average clustering coefficient (ACC) is defined as the fraction of a node's neighbors that are also connected to each other (taking the global average across all nodes). Intuitively, a higher ACC implies tight-knit communities and likely more triangles, while a lower ACC means the graph is more hierarchical with chain/tree-like structures. Another interesting characteristic is the Gini coefficient for node degree values, as most graphs have a moderate inequality/imbalance in node

degree values, but wiki-Vote happens to have a much larger value (also reflected in the significant difference between its mean and median). Lastly, our selected datasets possess different values for the Jaccard similarity between the top 20% of nodes in terms of degree centrality and the top 20% of nodes in terms of eigenvector centrality. wiki-Vote and email-Eu-core, for instance, seem to have a high overlap between which nodes both have a higher quality and quantity of connections, while facebook_combined and ca-GrQc have a significantly lower overlap, hinting that their graph structure may possess traits that contribute to some nodes with fewer connections having higher quality connections.

Examining which graphs exhibit ECD under different diffusion models, we notice some potentially noteworthy patterns. We first see that LAB led to an improved ratio compared to IC on every dataset except ca-GrQc, which may be due to the high fraction of closed triangles, as with limited attempts to activate neighbors, such a clustered environment makes it more likely that sampled neighbors have already been activated, preventing diffusion from spreading very far. On a similar note, while LAB didn't guarantee ECD, we believe it's worth highlighting that the three graphs that did exhibit ECD under LAB were those with the three lowest ACC values. A lower ACC essentially implies that the graph doesn't possess as many local redundancies, meaning connections to more influential nodes are especially crucial to proliferating the diffusion process across the more branch-like structure and reaching more distant regions (as in, "who you know is less important than who they know"). Overall, this demonstrates the value in creating diffusion models that better reflect real-world dynamics by modifying the assumptions of previously established theoretical models, as well as the importance of understanding the graph structure if possible when approaching the optimal seeding/influence maximization problem.

LABC led to ECD on all 5 of our datasets, though this shouldn't come as a surprise considering how we've defined the convincingness parameter. As our simulations were wrapping up, we realized that we'd essentially cheated in how we designed LABC as an extension to LAB by defining the probability of a given node successfully activating one of its neighbors in each independent attempt to be a function of the node's eigenvector centrality in the first place. We still believe LABC is worth discussing, as while it may not be the perfect diffusion model to capture ECD in a novel and insightful way, it still demonstrates the value of thinking critically about how behavioral science and sociology can be leveraged to design theoretical methods and models that better capture the intuitive aspects of the real world.

As another broader reflection, even the classical models of diffusion had some noteworthy quirks. LT didn't exhibit ECD for any of the datasets, perhaps a reflection of the fact that LT depends on the extent to which those around you have activated in coordination. In other words, the benefit of being connected to important neighbors is mitigated by the fact that a certain threshold of them would need to have been activated in order for diffusion to continue spreading. We also observed that IC generally yielded significantly high spread overall than LT, LAB, and LABC. Once again, this reminds us how subtle design choices in how we decide to model the world can have major consequences in the outcomes of experiments using those models. IC is the least restrictive of the models we examined for this project, as not only does it involve attempting to activate every neighbor, but it does so with a homogeneous probability for each independent attempt. Crucially, decisions regarding parameters (such as the probability of success in IC) significantly alter the final downstream outcome of the diffusion process. Better understanding the various theoretical tools we possess and improving our grasp at applying them correctly helps us maximize the benefit of leveraging theoretical models in the first place.

7 Conclusion

In this project, we proposed two new models of diffusion (LAB and LABC) to better reflect the fact that real-world interactions often are subject to limited attention/resource constraints. We compared these new models to the classical IC and LT models of diffusion, in hopes of understanding what diffusion conditions lead ECD to reflect empirical findings such as those from Banerjee et al. [2013]. Lastly, we concluded that graph structure plays a significant role in determining the final spread, often in cahoots with the specifics of a diffusion rule. We demonstrated this by analyzing some toy models of special graphs as well as running extensive simulations on real-world data, which hinted (among other things) that graphs with particular features such as lower ACC may more frequently exhibit ECD, especially when using diffusion models such as LAB and LABC that are designed to emphasize quality over quantity of connections.

As far as future directions, one difficult but important direction would be to do additional analysis on general and/or random graphs to reach more definitive theoretical conclusions. Experimental design on the simulation end can also always be improved, as a more effectively designed simulation script can lead to significantly improved runtime, allowing us to test on much larger datasets to see if our observations hold at scale. Equipped with an optimized script, another important layer to explore would be the effect of altering various parameters (e.g. μ, β, k) to see how they impact diffusion outcomes. Ona final note, we believe it'd be worthwhile to explore other extensions to LAB/LABC, especially determining a more robust formulation of convincingness to include in additional simulations.

References

Mohammad Akbarpour, Suraj Malladi, and Amin Saberi. Just a few seeds more: The inflated value of network data for diffusion. Working paper, 2021.

Sinan Aral and Marshall Van Alstyne. The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171, 2011. doi: 10.1086/661238.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. Using gossips to spread information: Theory and evidence from two randomized controlled trials. *The Review of Economic Studies*, 86(6):2453–2490, 2019.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

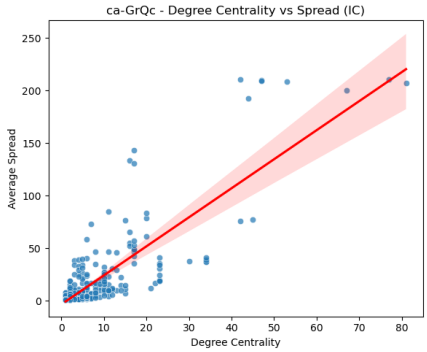
Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 1029–1038, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300551. doi: 10.1145/1835804.1835934. URL `https://doi.org/10.1145/1835804.1835934`.

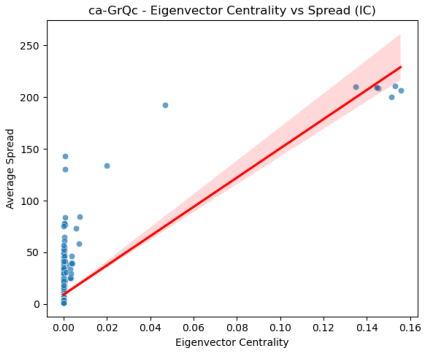
Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope, 2008. URL `https://arxiv.org/abs/0812.1045`.

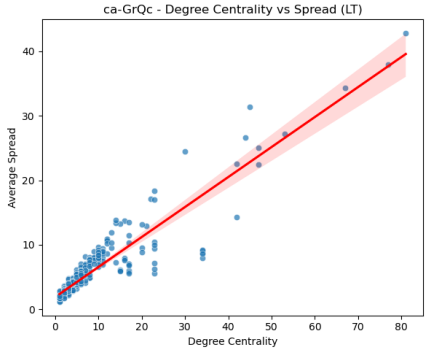
D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. *KDD*, 137:146, 2003.

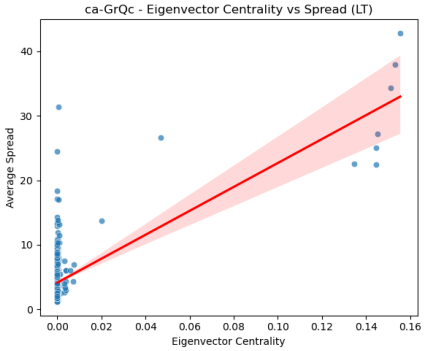
Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Competition among memes in a world with limited attention. *Scientific Reports*, 2:335, 2012. doi: 10.1038/srep00335.

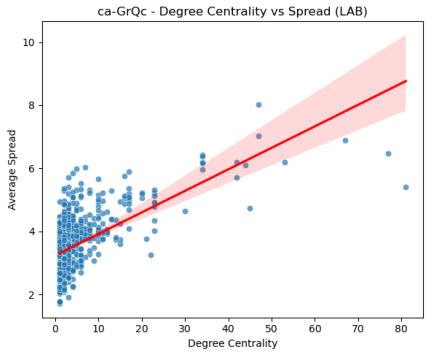
A Additional Plots

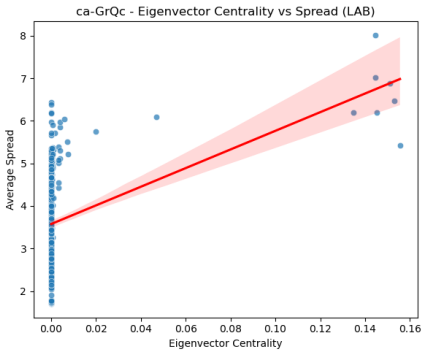


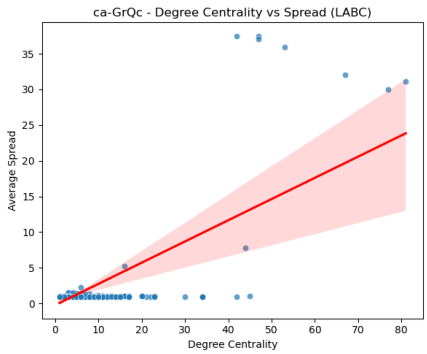












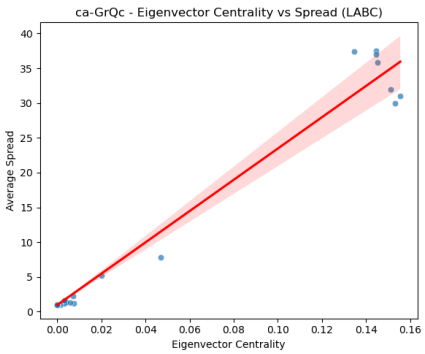


Figure 7: ca-GrQc Scatterplots

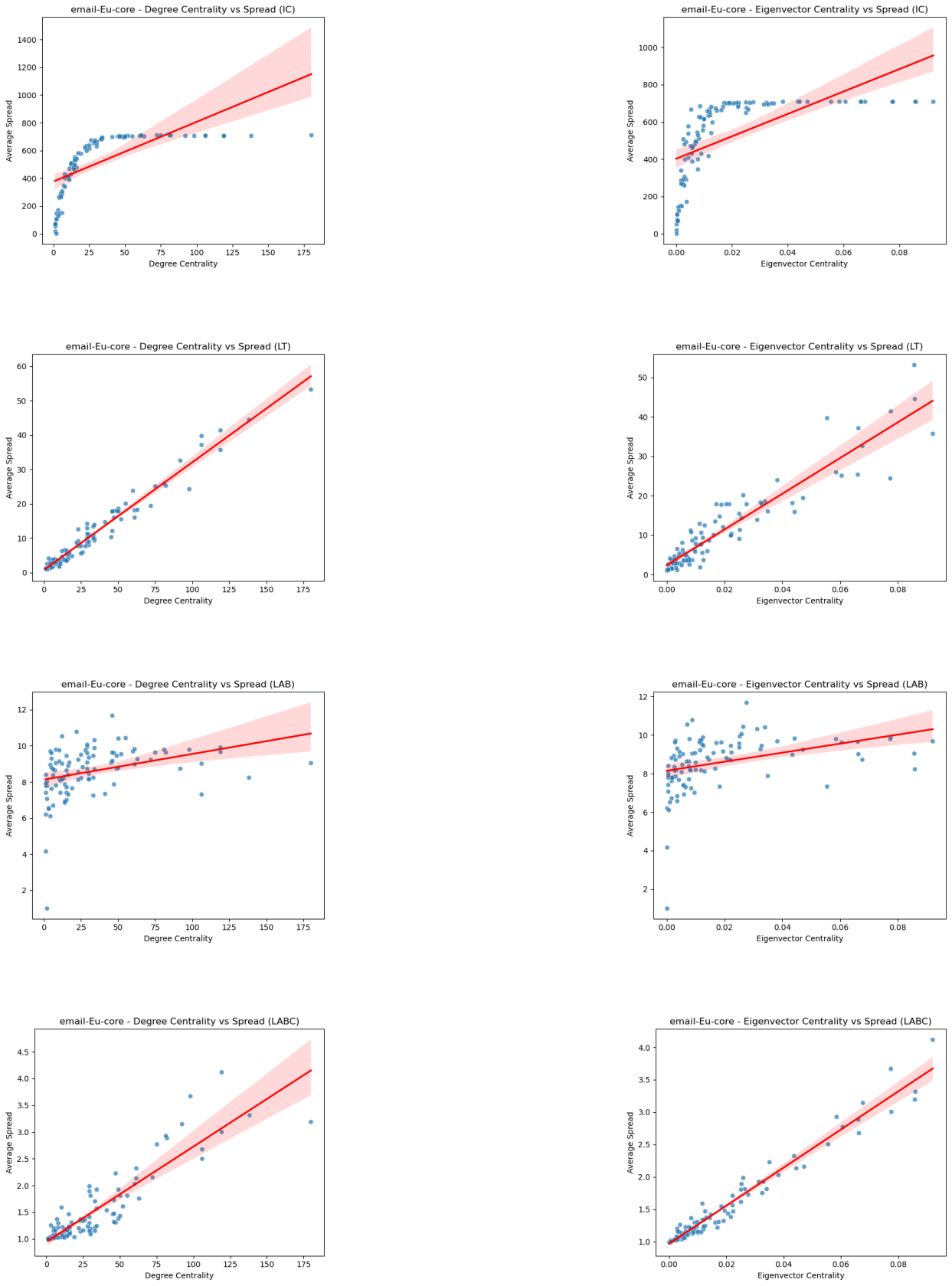


Figure 8: email-Eu-core Scatterplots

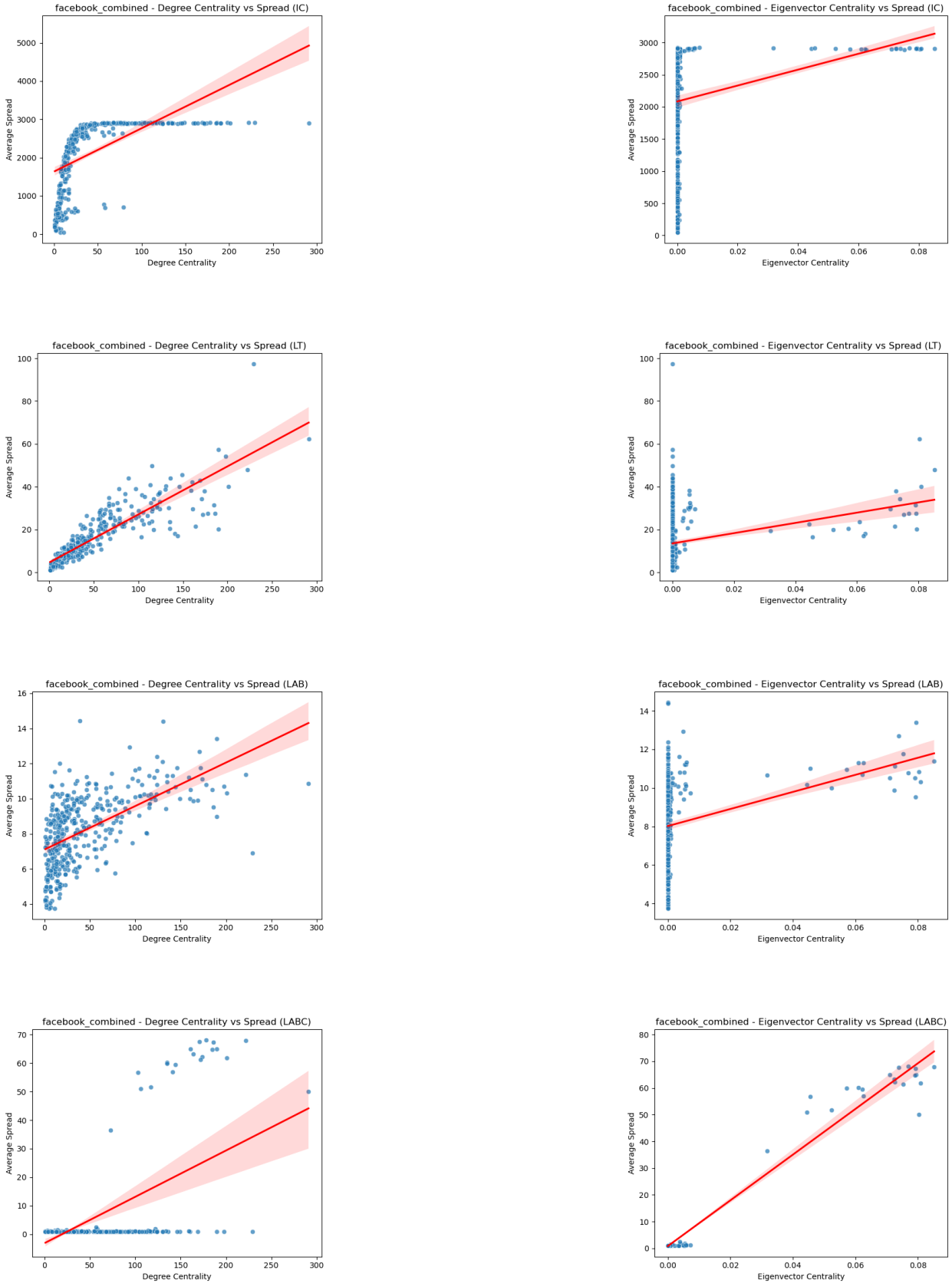


Figure 9: facebook_combined Scatterplots

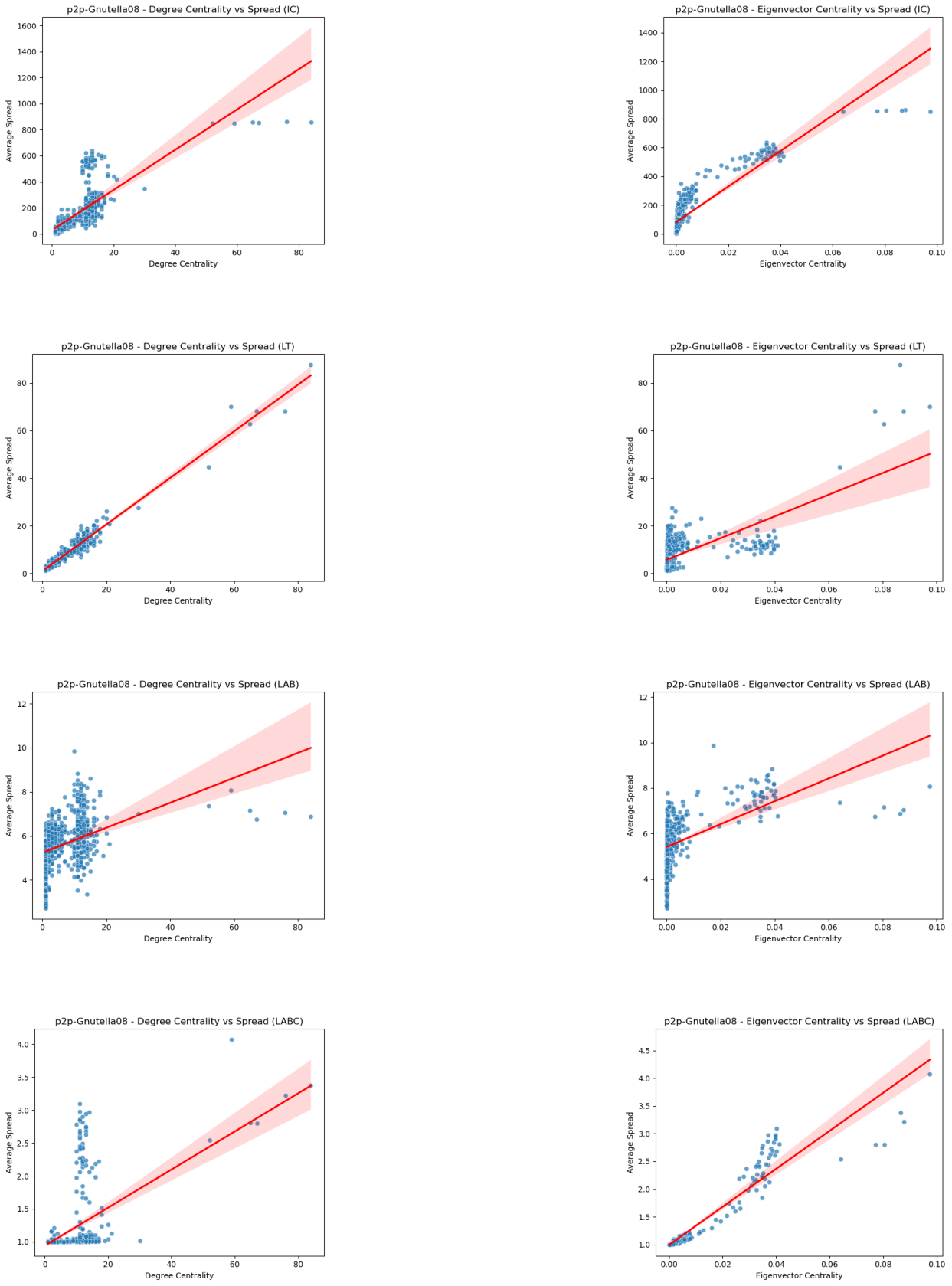


Figure 10: p2p-Gnutella08 Scatterplots

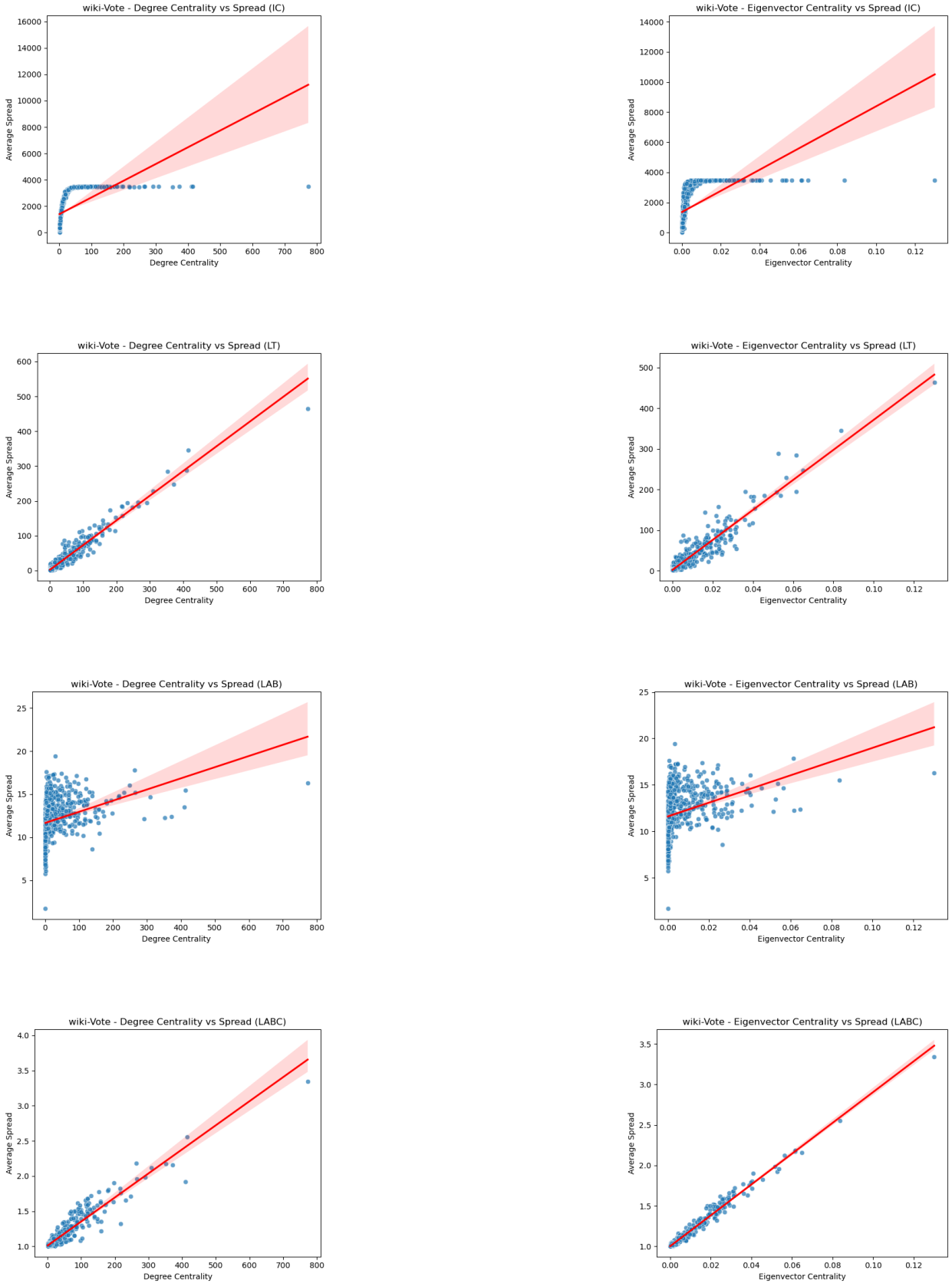


Figure 11: wiki-Vote Scatterplots