Chapter 2: CUPID

*Introduction / Literature Review*

Measuring New and Expanded Opportunities

Cumulative advantage is a key driver for the development of scientific stars (Mali et al. 2012, 235), a term that refers to a specific network structure. Networks consist of actors (researchers) and the relationships among them (ties) (Mali et al. 2012, 216). Social Network Analysis (SNA) provides a framework for understanding these structures, focusing on the relationships among actors within a network (Borgatti et al. 2022, 2; Mali et al. 2012, 216). The modular structure of researcher networks operates across disciplinary, sectoral, and geographical boundaries (Mali et al. 2012, 219; Vacca et al. 2015). These actors, or nodes, can be characterized by various categorical attributes, such as department affiliation, or continuous, like years of geographical distances (Mali et al. 2012, 219). The relationship in this context, termed as ties or edges, connects researchers to each other and can be quantified in multiple ways, including the frequency of interactions over a given period (Borgatti et al. 2022, 2; Mali et al. 2012, 216).

For example, Newman (2001) undertakes a comprehensive study of social networks, specifically focusing on scientific collaborations. Newman leverages the co-authorship of scientific papers as an unbiased and scalable measure for mapping social connections within the scientific community. Gathered data from multiple scientific databases, such as MEDLINE and the Los Alamos e-Print Archive, Newman (2001) created the network that tied each researcher in the network to all other researchers with whom they co-authored a paper within a five-year window (1995-1999). These ties interlink through common nodes, forming paths and, ultimately, a network. Within this network, frequently interacting actors may form a distinct subgroup (Borgatti et al. 2022, 2). Newman (2001) found that researchers tend to collaborate with peers who have gained influence through numerous prior joint projects, following a pattern of preferential attachment.

Co-authorship is a common type of relationship used to study scientific collaboration. In their book chapter, Mali et al. (2012) explore the complexities of scientific collaboration using co-authorship networks for their example but highlight various other collaborative activities, such as shared editorship, joint supervision of research projects, collaborative research proposal writing, participation in formal research programs, and the organization of scientific conferences (Mali et al. 2012, p 213).

Historical grant proposal application data from 2016 and 2020 creates multiple networks, including five-year and yearly networks. Nodes are faculty who collaborate within the given time frame, and edges are formed when any two faculty co-propose. Another grant proposal within the bounds of the network links these faculty to other faculty, creating a co-occurrence network (Borgatti et al. 2022). Faculty who proposed alone are removed from the network because we are examining collaborative proposals. Faculty who did not propose within a single year are removed from the network during network modeling.

The grant proposal network does not reveal the true social relationship between the faculty. While some faculty take on the role of PI on grant proposals, the reasoning for the role varies. Because of this, all individuals who share a grant together are considered equal, with no particular direction that connects the nodes.

Mali et al. highlight the foundational elements of modern social network analysis (SNA) as identified by Freeman (2004): a focus on structural analysis of actors within social relations, the use of systematic empirical data, extensive use of graphical imagery, and a foundation in formal, mathematical, and computational models (Mali et al. 2012, p 216). By leveraging SNA, I analyze the web of grant proposal collaboration, indicating how relationships and network structures contribute to developing scientific work. Using network visualizations, node and network metrics, and exponential random graph models, I describe BSU’s grant proposal collaboration networks and how they evolve between 2016 and 2020.

Methods

For example, the local property of a node in the network is **degree centrality**, deﬁned as the number of ties a node has (Mali et al. 2012, 214; Borgatti et al. 2022, 171). A high degree centrality takes the shape of a star, where one node has many ties to other nodes compared to most other nodes in the network. Its interpretation can vary based on the nature of these ties (Borgatti et al. 2022, 172). A star structure in team science networks may indicate a significant inequality in collaborative offers, as few scientists or scholars receive disproportionate offers to collaborate (Moody 2004). The cumulative advantage in science posits that scientist already recognized for their contributions are more likely to gain further recognition and resources (Mali et al. 2012, 235). This concept, drawing parallels to the biblical passage in Matthew's Gospel and referred to as "The Matthew Effect," implies a disparity in the distribution of resources and opportunities within the scientific community, where established researchers gain disproportionately more funding and power while emerging scientists face challenges in achieving recognition and success (Mali et al. 2012, 235–36). This concept highlights how normal social behaviors can thwart the GCs’ investment goal to expand research opportunities across campus.

Networks formed through this *preferential attachment* suggest a scale-free structure characterized by a power-law degree distribution where burgeoning scientists tend to collaborate with established ‘scientific stars’, reflecting the principle of cumulative advantage in science (Mali et al. 2012, 215; Vacca et al. 2015). This scale-free structure could indicate a hierarchical network dominated by a few highly connected individuals or "hubs" (Mali et al. 2012, 236). One method I use to examine the presence of scientific stars is the **degree distribution**. A declining degree distribution indicates that most network members have few ties, and few members have many ties (Harris 2014, p. 17).

**Betweenness centrality** measures a node’s frequency along the shortest paths between other node pairs (Borgatti et al. 2022, 182). It is interpreted as a node’s potential to control or regulate the flow through the network, playing a gatekeeper or broker role (Borgatti et al. 2022, 183). With their control over resources and opportunities, gatekeepers play a crucial role in shaping the network's topology (Mali et al. 2012, p 236). I examine the **betweenness distribution** to identify this phenomenon.

The **connectedness** score illuminates the level of structural cohesion (Borgatti et al. 2022, 201–3). Comparing the connectedness across each year’s network depicts the change in structural cohesion of the grant proposal network over time. Analysis from degree distribution and connectedness could be used to intentionally connect researchers across diverse modules, such as spanning structural holes and counterbalancing preferential attachment, as Vacca et al. showcase. The potential of network interventions is to overcome inherent biases in collaboration patterns and to bridge gaps between disparate scientific communities. Vacca et al.’s (2015) approach offers a pragmatic pathway for fostering cross-disciplinary team science and enhancing the cohesion and diversity of scientific research networks.

I use Bolger's (2021) degree of interdisciplinary research by evaluating disciplinary distance in the grant proposal network (see introduction). I evaluate co-grant proposals across distinct super-disciplines (e.g., an ecologist working with a social scientist) (Bolger 2021). This categorization distinguishes collaborations spanning 'hard' sciences (natural and applied sciences) and 'soft' sciences (social sciences and humanities), offering a more granular understanding of interdisciplinary research dynamics (Bolger 2021).

The analysis of subgroups and the overall network structure allows for the examination of shared attributes, offering insights into the collaborative dynamics in scientific communities (Borgatti et al. 2022, 2–3, 214). Research specialties can be identified as a cluster of collaborating scientists responsible for producing a significant number of innovative concepts and ideas (Moody 2004; Vacca et al. 2015). Collaboration *within disciplines* often leads to the emergence of distinct clusters within research collaboration networks, indicative of a *small-world* network structure marked by high local clustering and minimal steps between clusters (Vacca et al. 2015; Mali et al. 2012, 215). This structure contrasts with a cohesive core, characterized by an increasing trend of authors from various disciplines collaborating with each other (Moody 2004). I use **network visualizations** showing the researcher’s affiliated **college attribute** to illuminate possible disciplinary and short-distance interdisciplinary clustering (Bolger 2021). The grant networks did not allow for short-distance examination because there are too many departments (93) to analyze effectively.

In addition to network visualizations, network statistics can illuminate interdisciplinary patterns. In small-world networks, there is a notable pattern of dense local connections among actors, yet these actors are separated by only a few intermediary steps (Moody 2004). This structure contrasts with a cohesive core, characterized by an increasing trend of authors from various disciplines collaborating with one another (Moody 2004). The **clustering coefficient**, a measure reflecting the network's tendency for triadic closure, is calculated by the ratio of the actual number of closed triangles to the potential number of triads that could possibly contain at least two ties (Goodreau et al.,2009). Interestingly, a network's propensity for clustering often corresponds with increased path lengths, suggesting that as clusters become more defined, the distance between separate clusters can grow (Borgatti et al. 2022, 199). Networks with clusters that cross super-disciplines form “invisible colleges” that drive the intellectual and creative output of the scientific community (Mali et al. 2012, p 236). The application of Exponential Random Graph Models (ERGMs), discussed below, allows for modeling this local clustering phenomenon within the network.

**Density** is a fundamental concept that offers insight into a network's overall structure and interconnectivity. Norton et al. define density as the “ratio of the number of actual links to the number of possible links in the network” (2017, 6). This ratio provides a quantitative measure of how interconnected the individuals within the network are. Borgatti et al. further explain that density indicates the likelihood of any two individuals within the network being connected (2022, 195–96).

Lusher, Koskinen, and Robins assert that the network structure is a product of the social process that produced it and cannot be assumed to be known a priori (2013, 41). This statement highlights that network density often results from the underlying social interactions and processes. In the framework of ERGMs, density is closely linked to the edges term.

ERGMs

The landscape of SNA has been profoundly transformed by the introduction of ERGMs (Mali et al. 2012, 218). ERGMs are a specific category of statistical models that articulate the likelihood distribution of network graphs, premised on the assumption that network connections form patterns or configurations that recur more frequently than chance would predict (Harris 2014, 33). These configurations vary broadly, offering adaptability for various contexts, with a positive parameter value indicating a configuration’s propensity to occur more often within the network data (Caimo and Gollini 2020, 2).

The edges term in ERGMs resembles the intercept term in logistic regression models (Harris 2014, p. 52–53). The significance of this term reveals if the network's density deviates from a random network’s expected density (Harris 2014, p. 52–53). The edges term in an ERGM is translated into a probability, showing the likelihood of any two nodes in the network being connected (Harris 2014, p. 52–53). This probability, reflective of the network's density, indicates the extent to which ties in the network are not randomly formed.

The null model of an undirected network, described by Harris (2014, p. 39–47), only includes a single edges term representing the number of connections in the network. This model sets a baseline by capturing the network's overall propensity to form edges (its density) while disregarding other structural features. The statistical significance of the edges term in more complex ERGMs, implies that the network's structure is not random but is likely influenced by underlying principles (Harris 2014).

In developing the ERGMs for this study, I adopt a methodical, stepwise approach to integrate additional nodal attribute terms. To assess the fit of these progressively complex models, I utilize the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These criteria serve to evaluate model fit, balancing the deviance reduction against the complexity introduced by additional parameters, thereby penalizing over-parameterization (Harris 2014, 63). Terms that contribute to a reduction in AIC or BIC values are maintained in subsequent model iterations. The construction of these models follows a deliberate sequence, aligned with the order of term introduction as detailed below. This sequencing ensures that the analysis prioritizes terms of paramount relevance to the objectives of this study.

Dyadic Independence Terms

As by Lusher, Koskinen, and Robins (2013, 22) describe, local network configurations are hierarchically nested structures ranging from dyads, formed by a single tie between two nodes, to more complex formations like stars and triads. In this framework, interaction terms for nodal attributes, as Harris (2014) noted, play a crucial role, especially in analyzing dyadic relationships where the attributes of both nodes are considered. Among these, homophily, the tendency of nodes sharing an attribute to form connections, is a prevalent concept.

Goodreau et al. (2009) examine friendship networks, demonstrating the use of several ERGM terms. They explain *selective mixing* as the propensity of individuals to form connections based on shared attributes (Goodreau et al. 2009). They define **uniform homophily** as the tendency to form ties with others who have similar (homophily) or different (heterophily) attributes (Goodreau et al. 2009).

To evaluate the grant proposal network for long-distance interdisciplinary collaboration (Bolger 2021), I first investigate uniform homophily using the `*nodematch*` term. The `*diff = TRUE*` argument separates parameters for each college category, allowing the model to capture the propensity for faculty within the same college to co-propose more than would be expected by chance. Each college’s statistics quantify the extent of collaborative grant proposal engagement among faculty within the same college. An affirmative coefficient indicates homophily, signifying discipline-centric or “short-distance” interdisciplinary collaborations. Conversely, a negative coefficient implies heterophily, suggesting “long-distance” interdisciplinary collaborations. The working hypothesis posits that faculty members tend to co-propose with other faculty members within their own college.

Another selective mixing term that Goodreau et al. (2009) describe is **differential homophily**, a propensity to form ties specific to individual categories. The likelihood of forming a tie depends on a particular attribute that differs across various categories of that attribute. As an illustration, Lane et al. (2020), investigated the use of Evidence-based Instructional Practices (EBIPs) in college STEM courses, examining the communication ties relative to EBIP usage among instructors. Their findings revealed distinct interaction patterns based on EBIP familiarity, suggesting a nuanced picture of knowledge diffusion (Lane et al. 2020). Similarly, my study assigns a quartile attribute based on co-proposal counts to explore differential homophily within the grant proposal network.

I aim to determine if researchers who frequently co-propose grants tend to collaborate with others with similar co-proposal activity. I explore the concept of differential homophily within these quartiles using the `*nodemix*` term. This approach allows me to examine whether there is a tendency for high proposers to collaborate with other high proposers, which could indicate a "rich getting richer" phenomenon. Conversely, if high proposers frequently collaborate with low proposers, this might suggest a mentorship dynamic. If the implications of these patterns are significant, they potentially indicate disparities in resource distribution and opportunities within the scientific community.

Dyadic dependence terms

Transitioning from this exploration of selective mixing, the study also considers the influence of individual characteristics on the propensity to form collaborative ties. **Sociality** captures individuals’ intrinsic tendencies to form friendships (Goodreau et al. 2009). It is influenced by various factors such as personality, sociodemographic characteristics, or even external circumstances. Goodreau et al. considered sociality a social process contributing to the outcome, degree. I examine the effect of a faculty member’s college on their propensity to co-propose using the term `*nodefactor*`. Each college has statistics, effectively measuring how much more or less likely faculty in the specified college are to co-propose on a grant compared to the reference college.

I complement the AIC and BIC model fit assessments with a comparison of network characteristics between the observed data and simulated networks, following the methods by Harris (2014, 63–70). This comparison reveals a notable misalignment in the degree distribution and the distribution of edgewise shared partners, underscoring the necessity for incorporating dyadic dependence terms.

Dyadic independence ERGMs, which include only nodal attribute terms akin to traditional logistic regression, postulate that the probability of a tie is contingent solely upon the attributes of the actors involved, with tie values being mutually exclusive (Goodreau et al., 2009). This is congruent with maximum pseudolikelihood estimation (MPLE) mirroring maximum likelihood estimation (Goodreau et al. 2009).

Nevertheless, such conventional statistical models presuppose the independence of observations, a notion at odds with the complexities of human social behavior, which is multifaceted and intention-driven (Lusher, Koskinen, and Robins 2013). Certain network patterns are important based on specific social science theories, adopting a particular dependence hypothesis and definition for local configurations (Lusher, Koskinen, and Robins 2013, 19). For instance, the likelihood of a tie forming between two individuals can depend on whether they share common ties in the network, reflecting a tendency for triadic closure (Lusher, Koskinen, and Robins 2013, 69–71).

Geometrically weighted terms in ERGMs capture the complexity of social networks by accounting for the dependency structure among ties. These geometric terms reflect high-order dependencies, introducing challenges in estimating model parameters (Lusher, Koskinen, and Robins 2013, 69–71; Hunter, Goodreau, and Handcock 2008). Models with geometrically weighted terms require Markov chain Monte Carlo (MCMC) simulation methods to address model degeneracy (Lusher, Koskinen, and Robins 2013, 71; Hunter, Goodreau, and Handcock 2008, 254). MCMC works by generating a sample of possible networks that could theoretically have generated the observed data, allowing for the estimation of parameters that best represent the underlying social processes shaping the network (Harris 2014, 71; Hunter, Goodreau, and Handcock 2008, 254).

The **geometrically weighted edgewise shared partners** (GWESP) and geometrically weighted dyadwise shared partners (GWDSP) terms capture the concept of transitivity in network structures. GWESP how the presence of shared partners between two individuals influences the formation of new ties (Goodreau et al., 2009; Lusher, Koskinen, and Robins 2013, 69-71). Unlike simple triad closure, which might occur through incidental contact, transitivity reflects a deeper process where shared friends or collaborators lead to direct connections based on perceived social value or affinity (Goodreau et al. 2009). By incorporating the GWESP term into the models, I quantitatively assess the network’s clustering by considering how much an existing shared co-proposal contributes to forming additional co-proposers.

As Harris (2014, 85) explains, a statistically significant GWESP coefficient implies that the likelihood of tie formation between two individuals is higher than expected by chance, given all other factors are held constant. In other words, shared partners significantly increase the chances of two faculty members collaborating on a grant proposal. If the GWESP coefficient were negative, it would suggest a network where shared partnerships are less likely to lead to new ties, possibly indicating a network less driven by collaborative clusters (Harris 2014, 85). In the context of the grant proposal network, a significant positive GWESP coefficient would support the idea that faculty are more likely to co-propose with others who have mutual collaborators, reflecting a tightly knit community where collaboration is fostered through established connections (Harris 2014, 85). This pattern is characteristic of networks where knowledge and resources are often exchanged within well-defined local clusters, indicating disciplinary research or thematic communities (Mali et al. 2012, 236).

Building on my investigation of cumulative advantage, I investigate the network's **geometrically weighted degree** (GWD). GWD is integral for modeling the degree distribution within networks where the presence of higher-degree nodes is given more weight, indicating a network with a greater number of highly connected nodes (Harris 2014, 83).

A node with two ties is a 2-star node, and a node with k ties forms a k-star. Alternating star parameters, or geometrically weighted degree parameters, are used to model the distribution of nodes with varying numbers of ties (Lusher, Koskinen, and Robins 2013, 65–66). These parameters apply weights with alternating signs to different star counts, which regulate the impact of nodes with numerous connections, mitigating abrupt transitions in network density (Lusher, Koskinen, and Robins 2013, 65–66). When significant, these terms indicate that the network structure cannot be dismissed as random; rather, it is shaped by underlying social processes (Hunter, Goodreau, and Handcock 2008).

A significant positive coefficient for a GWD term in an ERGM suggests that the network is more likely to exhibit nodes with higher degrees than would be expected by chance (Harris 2014, 85). This could imply a tendency towards preferential attachment (Harris 2014, 85), where certain nodes act as hubs within the network (Mali et al. 2012, 236). Conversely, a significant negative coefficient would suggest an inclination against such hubs, indicating a more uniform or egalitarian distribution of ties across nodes (Harris 2014, 85). However, the nuances of these coefficients should be interpreted with caution due to the intricate way a single tie can affect the overall shared partner distribution within the network (Harris 2014, 85).

As recommended by Harris (2014, 72), I specify the number of networks to be sampled (MCMC sample size = 10,000), the number of networks to be ignored at the beginning of the sample (MCMC burn-in = 100,000), and the number of networks to skip over between sample networks (MCMC interval = 1,000) to reduce the likelihood of model non-convergence. I set the seed to ensure that model estimates are replicated. Finally, I fix the decay parameter (α), starting at 0.1, and increasing until the log-likelihood ceases to improve the AIC and BIC values (Harris 2014, 72). The final model for the 5-year network has a decay parameter of 0.3 for both geometrically weighted terms.

This concept highlights how normal social behaviors can thwart the GCs’ investment goal to expand research opportunities across campus.

Networks formed through this *preferential attachment* suggest a scale-free structure characterized

Goodness of Fit

Goodreau et al. (2009) discuss how the GWESP statistic incorporates a decreasing marginal effect on the probability of tie formation between two actors as the number of shared partners grows. This reflects the understanding that while initial shared connections significantly impact forming a tie, each additional shared connection beyond the first few contributes less to the likelihood of new tie formation.

\_\_

WORK HERE NVIVO metrics ERGMs CUPID Network Stats and up

In their study, Vacca et al. (2015) highlight the significance of recognizing and utilizing the modular structures within a university’s scientific collaboration networks to facilitate effective network interventions. They demonstrate that mapping co-authorship and grant proposal networks, identifying unconnected researcher groups, and employing strategic interventions can significantly enhance the network's overall structure, promoting interdisciplinary collaboration and innovation.

Mali et al. (2012) also acknowledges the complexities and challenges in fostering and measuring interdisciplinary research due to its demands for extensive networks, time, and mobility among researchers (Mali et al. 2012, p 222).

Vacca et al. not only mapped the university’s scientific collaboration network in their 2015 study but also surveyed the peripheral researchers receiving network treatments. The researchers discovered that familiarity with each other's work decreases as researchers are further apart in the network, implying that network data can effectively map a university's research activities (Vacca et al. 2015). Additionally, they observed that researchers are more skeptical about collaborating with more distant individuals in the network (Vacca et al. 2015). This skepticism is particularly pronounced in the context of grants and patents, potentially due to trust issues (Vacca et al. 2015). They recommend adding incentives for collaboration to motivate distant collaborations (Vacca et al. 2015).

The findings of Vacca et al. (2015) underscore the utility of employing methods like surveys to glean insights into the dynamics of scientific collaboration networks.

Limitations:

There is a limitation to Goodreau et al. interpretation of sociality contributing to the degree value. While degree is a directly measurable attribute, sociality is more of an inferred characteristic based on observed patterns of tie formation. For instance, if a person has a high degree, it could be inferred that they have high sociality. However, this is not always a direct correlation, as other factors like organizational structure, external incentives, or opportunities for interaction can also influence the degree. For example, workload policy and grant funding need differ across academic departments.

WORK HERE: GINI COEFFICIENT

*\*Remember to include all nodes in the degree centrality table used to calculate the Gini*