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).

THE GRANT NETWORK methods

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.

WORK HERE: GINI COEFFICIENT

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

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, p 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, only includes a single edges term representing the number of connections in the network (2014, p. 39–47). 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).

Goodreau et al. (2009) apply ERGMs to analyze adolescent social networks, focusing on the concept of **sociality**. This concept captures individuals’ intrinsic tendencies to form friendships. 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 for ties 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.

Local configurations are nested where a single tie between two nodes forms a dyad, a node with two ties is a 2-star, and three nodes that are tied together form a triad (Lusher, Koskinen, and Robins 2013, 22). Interaction terms for nodal attributes account for the attributes of both members of a dyad (Harris 2014). Homophily is the most commonly used interaction term; two nodes sharing an attribute have a greater propensity to form relationships (Harris 2014). Conversely, heterophily is where two nodes of different attributes are more likely to form relationships (Harris 2014).

Goodreau et al. (2009) examine *selective mixing*, the propensity of individuals to form connections based on shared attributes. 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 investigate uniform homophily using the `nodematch` term. Each college’s statistics reflect the number of co-proposals between faculty members sharing the same college. I hypothesize that faculty co-propose with other faculty members within the same college. Goodreau et al. (2009) define **differential homophily** as a selective mixing process that varies by attribute categories, meaning that the propensity to form ties is specific to individual categories. The likelihood of forming a tie depends on sharing a particular attribute, but the propensity can differ across various categories of that attribute.

An example of a differential homophily study is Lane et al. (2020), who explored the utilization of Evidence-based Instructional Practices (EBIPs) in college STEM courses. They study how the frequency of EBIPs use by instructors at academic institutions predicts ties in a communication network (Lane et al. 2020). Specifically, they examine whether educators who frequently use EBIPs engage in discussions with those less familiar with them, a factor critical for the diffusion of EBIPs across educational settings (Lane et al. 2020). The study utilized a Guttman scaling survey to assess faculty members' knowledge and use of EBIPs, ranging from awareness to regular application (Lane et al. 2020). These levels were then transformed into ranked quartiles assigned as node attributes in their communication network (Lane et al. 2020). Using ERGMs, the researchers analyzed the likelihood of teaching discussion ties predicting EBIP usage discussions (Lane et al. 2020). The results indicated that low EBIP users were least likely to be cited as discussion partners by both high and low EBIP users (Lane et al. 2020). This lack of interaction suggests that the knowledge of EBIPs is unlikely to reach the less experienced educators through secondary diffusion processes (Lane et al. 2020).

I mirror this approach by assigning a quartile attribute based on the count of co-proposals. 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.

Standard statistical approaches assume independence of observations, but humans are intentional beings with multiple motivations for and expressions of social action (Lusher, Koskinen, and Robins 2013). Dyadic independence ERGMs only contain node attribute terms and are very similar to traditional logistic models (Goodreau et al., 2009). The probability of a tie does not depend on the value of other ties, only the actors involved. Maximum pseudolikelihood estimation (MPLE) is equivalent to maximum likelihood estimation (Goodreau et al. 2009).

Dependency

ERGMs can address the issue of interdependence by incorporating dependency between network ties, which permits understanding the complex combination of social processes by which network ties are formed. 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).

One of the critical concepts in understanding dependency in networks is triad closure. This concept suggests that if two people have a mutual friend, they will likely become friends (Goodreau et al., 2009). This concept is colloquially understood as “a friend of a friend is a friend” (Lusher, Koskinen, and Robins 2013, 62). Triad closure should not be confused with transitivity, which has its basis in structural balance theory (Goodreau et al. 2009). While triad closure is where two people encounter each other through the shared time with the third person, **transitivity** involves a more cognitive dimension—where a social value or affinity between two individuals leads them to form a tie directly with each other (Goodreau et al. 2009).

I incorporate geometrically weighted edgewise shared partner (GWESP) distribution to quantify the effect of transitivity within the grant proposal network. The GWESP term in ERGMs quantifies the extent of local clustering by considering how shared partnerships contribute to the formation of ties (Lusher, Koskinen, and Robins 2013, 69–71).

The GWESP statistic models the impact of each additional shared partner on the probability of tie formation, with the effect diminishing as the number of shared partners increases (Goodreau et al. 2009). This diminishing return captures the complex interplay of social processes in network tie formation, where each new shared connection contributes less to the likelihood of a new tie as the number of shared connections grows (Goodreau et al. 2009). GWESP is the alternating triangle statistic that represents

(Lusher, Koskinen, and Robins 2013, 71)

Models with the GWESP term require Markov chain Monte Carlo (MCMC) simulation methods to address model degeneracy, .

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. This could imply a tendency towards preferential attachment, where certain nodes act as hubs within the network (Harris 2014, 85). Conversely, a significant negative coefficient would suggest an inclination against such hubs, indicating a more uniform or egalitarian distribution of ties across nodes. 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).

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

\_\_\_

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.