Balancing the Pain and Gain of Hobnobbing

Utility-Based Network Building over Attributed Social Networks

Yijin Cai*
Beijing Lab of Intelligent Information Technology, School of Computer Science, BIT
Beijing, China
2120160980@bit.edu.cn

Hong Zheng
Beijing Lab of Intelligent Information Technology, School of Computer Science, BIT
Beijing, China
hongzheng@bit.edu.cn

Jiamou Liu†
The University of Auckland
Auckland, New Zealand
jiamou.liu@auckland.ac.nz

Bo Yan
Beijing Lab of Intelligent Information Technology, School of Computer Science, BIT
Beijing, China
yanbo@bit.edu.cn

Hongyi Su
Beijing Lab of Intelligent Information Technology, School of Computer Science, BIT
Beijing, China
henrysu@bit.edu.cn

Yiping Liu
Beijing Lab of Intelligent Information Technology, School of Computer Science, BIT
Beijing, China
2120161088@bit.edu.cn

ABSTRACT

The establishment of interpersonal ties is a pivotal problem in the structural analysis of social networks. In particular, link recommendation problem asks for valuable future links to establish by an individual. Existing methods for this problem rely on link prediction that evaluates the likelihood of successful tie creation between two individuals. Such methods do not consider the social capital gained by agents, nor do they concern with the required cost of this process. In light of this limitation, we propose a utility-based network building problem, with an aim to strike a balance between the gained social capital – in the form of closeness centrality – and the cost of establishing ties. We propose algorithms to solve this problem over networks whose nodes may or may not be labelled with attributes, and test their performance on a range of synthesized and real-world social networks. By having multiple agents adopting utility-based network building strategies, we propose a suite of models of network formation and demonstrate empirically that the they capture important structural properties. In particular, we investigate the emergence of a core/periphery structure as a joint result of preferential attachment and network building strategies.

KEYWORDS

Network formation; link recommendation; core/periphery structure; social capital; link prediction; centrality

ACM Reference Format:
Yijin Cai, Hong Zheng, Jiamou Liu, Bo Yan, Hongyi Su, and Yiping Liu. 2018. Balancing the Pain and Gain of Hobnobbing. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 9 pages.

*Corresponding author
†Corresponding author

1 INTRODUCTION

Social positions play a fundamental role in our daily lives as they enable us to define who we are and what we ought to do. The creation of interpersonal ties thus takes a pivotal role in social network analysis. Studies revolve around two interrelated questions. The first is link prediction, which amounts to evaluating the likelihood of creating new ties between two individuals; the second is network formation, which concerns with how ties emerge and affect network topology. By occupying an advantageous position in a social network, an individual may gain control over information flow and achieve higher influence and status. On the other hand, establishing and managing social relations costs efforts, time and resources. A natural challenge lies in striking a balance between the costs and gains of establishing interpersonal ties.

We take up this challenge by bringing gain and cost into the same picture and propose strategies for individuals to build relations in the hope to arrive at an advantageous position. As online social networks (OSNs) such as Facebook, Google+ and Twitter increasingly dominate our social activities, there is a prominent need for advanced decision supports that help users to navigate the interpersonal web and enhance their social surroundings. In particular, a *link recommender system* evaluates links in the likely future and recommend top-ranked links to users. As an example, the "`People You May Know`" function in LinkedIn suggests to users new contacts to improve their career prospects [38].

Social capital refers to the intangible good brought to an individual from interpersonal interactions, either through being in a tightly connected clique that are bonded by kinship, support and trusts, or through accessing information flow [1]. An important form of social capital is *information brokerage*, i.e., network positions that bridge multiple closed groups of individuals and reach diverse population, which as argued by Uzzi and Dunlap in [41], bring private information, wide opportunities and power. The celebrated theory of R. Dunbar stipulates a cognitive upper bound on the amount of social contacts an individual may have [11]. A similar observation was made in management studies where every

person is bounded by "limited managerial attention", i.e., a finite capacity to exercise direct control of others [14]. Base on these theories, it is reasonable to assume that each individual has a finite budget which may be distributed among their social connections. The costs reflect the levels of difficulty or uncertainty involved in tie creation, and would certainly vary among different ties.

Undoubtedly gains and costs are two crucial factors that shape a person's decisions to form relations. Yet so far, very few existing works on link recommendation take these factors into account. A majority of existing works view link recommendation almost synonymously with link prediction [25]: Given a snapshot of the social network at a specific time, a link predictor attempts to approximate subsequent edge distribution as the network evolves [2, 15]. The highest predicted edges will then be recommended to the user.

Closeness centrality is widely-used for measuring the relative importance of nodes within a graph. It has been used to capture nodal importance since 1950s by Bevelas and Sabidussi, e.g., to identify central metabolites of a metabolic network [30] , to capture researcher impacts in co-authorship network [46], to analyze performance of business leaders [39], etc. Closeness is important also to game-theoretical models of multiagent systems, e.g., Branzei and Larson used it to study team formation [20].

Main contributions. (1) The primary contribution of the work lies in proposing the *utility-based network building problem*, which departs from existing schemes of link prediction and recommendation. By creating a tie, an individual gains in terms of centrality, but at the same time pays a cost that depends on link prediction: Higher probability of tie emergence means lower costs/uncertainty to establish this tie. (2) To solve this problem, we propose network building strategies for an individual and analyze their performance. These solutions also work on networks where agents are labeled with attributes. Empirically evaluations are performed on random network models and real-world social networks. (3) A further contribution is in the proposal of novel models of network formation. An enduring quest in complex networks seeks simple yet convincing formation models that give rise to real-life structures. Notable breakthroughs, e.g., Barabàsi-Albert scale-free and Watts-Strogatz small-world models, remain to have limitations. Here, through allowing all agents to practice network building strategies, we introduce a suite of new models, and show that their collective outcomes satisfy many desirable properties: power-law degree distribution, high clustering coefficient. In particular, we investigate the emergence of a core/periphery structure as a joint result of preferential attachment and network building strategies, thus filling a needed gap in the simulation of complex networks.

2 RELATED WORK

Common methods for link prediction exploit shared neighborhoods between nodes, and derives similarity e.g. common neighbors, Jaccard's coefficient, Adamic/Adar and resource allocation [24, 28]. A severe limitation exists in that node pairs with no common neighbor will be assigned zero probability. Random walk-based approaches, e.g., personalized PageRank, tackle this limitation and evaluate structural proximity between nodes [40]. We remark that a different notion of "utility-based link recommender" has been proposed in [23]. Contrary to our work, there utility is defined for the recommender, disregarding agents' costs and gains, and thus costs only occurs when a recommended link fails to form.

Social capital, proposed by J. Coleman [7], is widely believed to arise from the structure of interpersonal relations [26, 27]. In particular, distance measures how easy two nodes are reachable from each other via edges and constitutes a crucial indicator of social capital. Moskvina and Liu study a number of network integration problems that are NP-hard [31, 32]. Another natural idea is to maximize closeness centrality of a target node, where Yan et al. experiments on dynamic networks [44]. Crescenzi et. al. proves NP-hardness of the problem of maximizing centrality with $\leq k$ links [8]. Fink and Spoerhase study maximizing the collective betweenness centrality [13]. However, none of these works takes into account costs of establishing links. Leskovec et al. considers costs of monitoring nodes in a social network for spreading of information [21]. There, costs denote resources needed to watch over activities at particular nodes, but not on edges. To our knowledge, this current work is the first that correlates link cost negatively with nodal relatedness.

A range of random graph models exist to capture realistic network distributions, aiming to match such desirable properties as scale-freeness (power-law degree distribution), small-world property, community structure, etc. The classical Erdös-Rèny (ER) random graph model builds edges as independent Bernoulli random variables and results in a Poisson binomial degree distribution, with low clustering coefficient in general [12]. The most popular model for scale-free networks is Barabàsi-Albert (BA) model: The model creates edge in a preferential attachment scheme, where nodes with higher degree are more likely to be linked by new edges; These networks have in general low clustering coefficient [3]. Another well-used model is Watts-Strogatz (WS) model: The model starts from a regular ring lattice and randomly rewires edges with probability p. As edges are rewired, the network swings from a regular graph towards an ER random graph. When $p \in [0.01, 0.1]$, the graph typically exhibits high clustering coefficient, demonstrating a small world property. The model, however, fails to deliver power-law degree distribution [43]. Different to the stochastic models above, an economics-based paradigm views the individuals' behaviors in link formation as driven by utility; this approach offers insights into the cause of social and economical networks [4, 17].

Core/periphery structure is an important meso-level network property that is characterized by a highly connected core surrounded by a sparse periphery [16, 36, 45]. The property, firstly studied in a sociology [37], arises prevalently in collaboration networks [6, 42], financial networks [34], protein interaction networks [29], etc. Focus has been on detections of core/periphery structures [5] and numerous models have emerged, e.g., bow-tie, rich-club, and onion networks [9]. However, a general model that not only fits the structural metrics but also provides insight as to how this property emerges is still lacking. Through proposing utility-based network building strategies, we introduce models of network formation and show that they may give core/periphery structure if combined with preferential attachment.

3 UTILITY-BASED NETWORK BUILDING PROBLEM (U-NB)

3.1 Problem Setup

A *social network* is viewed as an undirected graph $G = (V, E)$ with no multiple edge nor self-loop. V is the set of *nodes* (or *agents*) in the social network and edges in E represent interpersonal ties. Two nodes connected by an edge $\{u, v\}$ is called *adjacent*; $N(v)$ denotes the set of all adjacent nodes to v and $N[v] \coloneqq N(v) \cup \{v\}$. A *path* is a sequence of nodes $v_1, \ldots, v_\ell$ where $\{v_i, v_{i+1}\} \in E$ for all $i = 1, \ldots, \ell - 1$. The *distance* $\mathrm{dist}_G(u, v)$ is the shortest length of a path between u and v if it exists, and is ∞ otherwise.

Utility. Social capital quantitatively evaluates structural advantage, which can be represented by *centrality*, i.e., how accessible a node is reachable from others. In particular, *closeness centrality* is affected by distance from a node to all other nodes, i.e., treating $1/\infty = 0$, for any $v \in V$, let

$$\mathrm{cl}_G(v) \coloneqq \sum \{\mathrm{dist}_G(u, v)^{-1} \mid u \neq v\}. \tag{1}$$

We view a node as an *agent* who aims to maximize its centrality through building relations with other nodes. Let $G = (V, E)$ be a graph, $S \subseteq V$ and $v \in V$ be a node. By $G \oplus_S v$ we denote the graph $(V, E \cup \{\{u, v\} \mid u \in S\})$ obtained by v establishing ties with all nodes in S.

Definition 3.1. Let $G = (V, E)$ be a graph and $v \in V$ be a *target*. The *utility* of a node $u \in V$ for the target v, $U_G(v, u)$, is defined as the improvement on the closeness centrality of v after creating edge $\{u, v\}$, i.e., $U_G(v, u) \coloneqq \mathrm{cl}_{G \oplus_{\{u\}} v}(v) - \mathrm{cl}_G(v)$. The utility of a set S is therefore

$$U_G(v, S) \coloneqq \mathrm{cl}_{G \oplus_S v}(v) - \mathrm{cl}_G(v). \tag{2}$$

Cost. Assume $u, v \in V$ are not already linked by an edge. Establishing a tie between u and v will clearly require less effort as u and v share more similar traits or social connections. Thus the cost is naturally affected by how likely a tie would emerge due to nodal similarity. Link prediction provides a way to approximate this cost. More formally, a link prediction algorithm outputs a function $\mathrm{score}\colon \{(v, u) \mid \{v, u\} \notin E, u \neq v\} \to \mathbb{R}$ where $\mathrm{score}(\{v, u\})$ denotes the estimated likelihood that the edge $\{v, u\}$ will emerge in G.

Definition 3.2. For $v, u \in V$ where $\{v, u\} \notin E$, we define the *link cost* $C_G(v, u)$ between v and u as

$$C_G(v, u) = \varphi^{-1}(\mathrm{score}(\{v, u\})) \tag{3}$$

where $\varphi\colon \mathbb{R} \to \mathbb{R}$ is a non-decreasing function. For $S \subseteq V$, define $C_G(v, S) = \sum_{u \in S} C_G(v, u)$.

The function φ is a parameter controlling the distribution of link costs and depends on the output $\mathrm{score}(v, u)$. In this work, we evaluate $\mathrm{score}(v, u)$ using *personalized PageRank* index, a well-established metric predicting likelihood of edges between any pairs of nodes [33]. The metric is an adaptation of PageRank: Given a starting node s, assign a score to every node u that captures the likelihood of a random walk from s (with restart) that reaches u [40]. More formally, let $\vec{a_u}$ be the column vector in the adjacency matrix of G corresponding to node u. The personalized PageRank vector $\vec{\mathrm{pr}}$ is defined by

$$\mathrm{pr}_u = (1 - \alpha) r_u + \alpha \, (\vec{\mathrm{pr}} \cdot \vec{a_u} / |N(u)|) \tag{4}$$

where $\alpha \in (0, 1)$ is a restart probability, and $r_u = 1$ if $u = s$ and $r_u = 0$ otherwise. pr_u is used as the link prediction score between s and u. Comparing to well-known link prediction methods such as those based on similarity indices, personalized PageRank is more suitable in our context as it gives a non-zero score for any reachable pairs of nodes, extending beyond pairs with nonempty overlapped neighborhoods. A limitation of this method is that a non-zero score is only given two pairs of nodes that are in the same connected components. If u, s are not connected by a path, pr_u is evaluated to 0, making the link cost between them infinite. To alleviate this problem, we introduce another parameter $\chi > 0$ and amend the cost function to the following *smoothed cost function* when the graph is not connected:

$$C_G(v, u) = (\chi + \varphi(\mathrm{score}(\{v, u\}))^{-1}. \tag{5}$$

Finally, we normalize $C_G(v, u)$ so that the maximum cost for (v, u) in the same connected component is 0.

Having defined the utility of agents and the costs of potential links, we formulate the *utility-based network building problem* (U-NB) as follows: The input to the problem consists of a network $G = (V, E)$, target $v \in V$ and a budget $B \in \mathbb{R}$. The output of the problem is a set $S \subseteq V$ such that v obtains the highest centrality in the network $G \oplus_S v$ while $C_{v,S}(G) \leq B$. More formally, the problem is defined as

$$\text{maximize} \quad U_G(v, S), \tag{6}$$

$$\text{subject to} \quad C_G(v, S) \leq B. \tag{7}$$

3.2 Attributed Networks

So far we have been focusing on social networks with only structural information. Our next goal is to extend the U-NB problem so that it applies also to attributed networks. An attributed network stores not only the interpersonal ties between agents, but also unary properties of agents. An example is an online social network where users has features such as age, gender, hobbies, professions, etc. Attribute information impacts significantly the likelihood of a tie establishing between individuals, and in general people with more common attributes would be more likely to be linked. Through incorporating personal attributes, a link predictor may provide more realistic output. For simplicity, we assume that all nodal attributes have discrete domain and thus the set of possible attributes is finite.

Definition 3.3. Let $x_1, x_2, \ldots, x_k$ be unary discrete features with finite domains $X_1, X_2, \ldots, X_k$, respectively. Let $X = \bigcup_{i=1}^{k} X_i$. An *attributed network* (with value set X) is a tuple (V, E, w) where (V, E) is a network and $w\colon V \to \prod_{i=1}^{k} X_i$ is an *attribute function*; $w(v)$ describes agent $v \in V$ where $w(v)[i] \in X_i$ is the ith feature of v.

A simple method to bring attributed network into the same framework as attribute-free network is through social attribute network [47]. The model augments a social network with attribute nodes. A *social attribute network* (SAN) contains two types of edges, the first being usual social ties between users and the second being ties between a user and an attribute. Weights on the edges control

the importance between attribute information and social ties. Gong then uses SAN to perform link prediction, emphasizing that the set of attributes should be pairwise distinct and attributes have different significance and obtain superior results than other link prediction methods over attributed networks [15].

Definition 3.4. For an attributed network (V, E, w) with attributes $x_1, \ldots, x_k$, the corresponding SAN is an edge-weighted graph $(V \cup X, E \cup E', g)$ such that the edge set $E' = \{\{v, w(v)[i]\} \mid v \in V, 1 \leq i \leq k\}$, and the edge weight function $g(e) = 1 - \lambda$ for $e \in E$ and $g(e) = \lambda$ for $e \in E'$, where $\lambda \in [0, 1]$ is an *edge weight index*.

Here we adopt SAN when evaluating link costs in attributed networks. The adjacency matrix of an SAN has $|V \cup X|$ rows and columns and entry-(i, j) is 1 if $\{i, j\} \in E$, λ if $\{i, j\} \in E'$ and 0 otherwise. The link prediction score $\text{score}(v, u)$ from v to u is defined as v's personalized PageRank pr_u in the corresponding SAN. This allows all notions in U-NB naturally carry over to attributed networks.

4 NETWORK BUILDING STRATEGIES

4.1 Three Heuristics

Our next goal is to design algorithms for U-NB for a given target v. We list some properties of the utility function U_G.

Theorem 4.1. *The utility U_G satisfies the following:*

(a) $U_G(v, \varnothing) = 0$.
(b) $U_G(v, S) \leq U_G(v, T)$ *for all* $S \subseteq T \subseteq V$.
(c) The function is submodular, *i.e., for any* $S \subseteq T \subseteq V$ *and* u *such that* $u \notin T$, $U_G(v, S \cup \{u\}) - U_G(S) \geq U_G(v, T \cup \{u\}) - U_G(T)$.

Proof. **(a)** is clear. For **(b)**, suppose $S \subseteq T \subseteq V$. For any node $u \in V$, $\text{dist}_{G \oplus_S v}(u, v) \geq \text{dist}_{G \oplus_T v}(u, v)$. Thus

$$\begin{aligned} \text{cl}_{G \oplus S}(v) &= \sum_{\text{dist}_G(u,v) < \infty} \frac{1}{\text{dist}_{G \oplus_S v}(u, v)} \\ &\leq \sum_{\text{dist}_G(u,v) < \infty} \frac{1}{\text{dist}_{G \oplus_T v}(u, v)} = \text{cl}_{G \oplus T}(v), \end{aligned}$$

implying that $U_G(v, S) \leq U_G(v, T)$.

For **(c)**, consider nodes $s \neq v$. Suppose the last edge of a shortest path from s to v in $G \oplus_{T \cup \{u\}} v$ is $\{u, v\}$ or belongs to $\{\{x, v\} \mid x \in S\}$. Then, $\text{dist}_{G \oplus_{T \cup \{u\}} v}(s, v) = \text{dist}_{G \oplus_{S \cup \{u\}} v}(s, v)$. Let $S_1 \subseteq V$ be the set of all such s, and let $S_2 = V \setminus (S_1 \cup \{v\})$. The last edge of any shortest path from $s \in S_2$ to v in $G \oplus_{T \cup \{u\}} v$ belongs to $\{\{x, v\} \mid x \in T \setminus S\}$. In this case, $\text{dist}_{G \oplus_{T \cup \{u\}} v}(s, v) = \text{dist}_{G \oplus_T v}(s, v)$.

Now we have

$$\begin{aligned} &U_G(v, T \cup \{u\}) - U_G(v, T) \\ &= \sum_{s \in S_1} \left(\frac{1}{\text{dist}_{G \oplus_{T \cup \{u\}} v}(s, v)} - \frac{1}{\text{dist}_{G \oplus_T v}(s, v)} \right) \\ &\qquad + \sum_{s \in S_2} \left(\frac{1}{\text{dist}_{G \oplus_{T \cup \{u\}} v}(s, v)} - \frac{1}{\text{dist}_{G \oplus_T v}(s, v)} \right) \\ &= \sum_{s \in S_1} \left(\frac{1}{\text{dist}_{G \oplus_{T \cup \{u\}} v}(s, v)} - \frac{1}{\text{dist}_{G \oplus_T v}(s, v)} \right) \\ &= \sum_{s \in S_1} \left(\frac{1}{\text{dist}_{G \oplus_{S \cup \{u\}} v}(s, v)} - \frac{1}{\text{dist}_{G \oplus_T v}(s, v)} \right) \\ &\leq \sum_{s \neq v} \left(\frac{1}{\text{dist}_{G \oplus_{S \cup \{u\}} v}(s, v)} - \frac{1}{\text{dist}_{G \oplus_S v}(s, v)} \right) \\ &= U_G(v, S \cup \{u\}) - U_G(v, S) \end{aligned}$$

Thus the function U_G is submodular.

□

Submodularity captures a "diminishing return" effect in many network-related processes such as information diffusion [18] and outbreak detection [21]. In general, finding optimal solutions with submodular objective function is NP-hard [19]. We thus focus on efficient heuristics to approximate optimal solutions of U-NB. we propose three greedy heuristics:

The first heuristic, Utility Strategy (SCLS), focuses on utility, i.e., it iteratively links v with $\arg\max\{\text{cl}_G(u) \mid u \in V \setminus S\}$. This simulates agents who prefer to build relations with those who have high status, disregarding cost. A downside of this strategy is that as costs of each link may be high, fewer number of ties will be established.

The second heuristic, Cost Strategy (SPP), builds ties that incur minimum costs, i.e., the strategy iteratively links v with $\arg\min\{C_G(v, u) \mid u \in V \setminus S\}$. This is the standard link prediction-based method where agents connect with those who are easy to establish ties with, ignoring utility.

The third heuristic, Unified Strategy (SUN), strikes a balance between utility and cost, iteratively building edges for v that maximize utility/cost ratio. The algorithm is presented in Alg. 1. Using a similar proof as in [19], one may obtain a performance guarantee of this strategy; see Thm. 4.2.

Theorem 4.2. *[19] The UnifiedStrategy in Alg. 1 achieves an approximation factor of* $\frac{1}{2} \times (1 - \frac{1}{e})$ *for the U-NB problem.*

4.2 Experiments on Network Building Strategies

We conduct experiments comparing the three network building strategies for the U-NB problem over both (attribute-free) synthetic networks and (attributed) real-world networks.

Data sets. We use random graph models, ER, BA, WS, as described in Sec. 2 and generate for each model graphs with 100, 200, and 500 nodes. We set parameters so that the average degree of each generated network is 4, as it occurs commonly in social networks.

Algorithm 1 UnifiedStrategy

Input: A graph $G = (V, E)$, target v, budget B
Output: set of nodes $S \subseteq V \setminus N[v]$ such that $C_{v,S}(G) \leq B$

```
Compute cost and utility for each node not adjacent to v
T ← ∅
while C_G(v, T) ≤ B do
    u_max := arg max_u {U_G(v, T ∪ {u}) − U_G(v, T)}
    T := T ∪ {u_max}
S := ∅
while C_G(v, S) ≤ B do
    u_max := arg max_u {(U_G(v, S ∪ {u}) − U_G(v, S)) / C_G(v, u)}
    S := S ∪ {u_max}
return S if U_G(v, S) > U_G(v, T); T otherwise.
```

E.g around 30 social networks on KONECT database have average degrees from 2 to 6. Also [43] uses an example with average degree 4 to introduce their model. We take DBLP data which contains co-authorship network of 2723 authors, captured from the DBLP Bibliography data. Each author has 43 distinct categorical attributes, denoting the number of the published researches. As the data set contains nine timestamps, we treat them as nine separate attributed networks [10].

Parameter calibration. We need to calibrate the following parameters: function φ and value χ in computing link costs, variables α, λ in the SAN model, and the total budget B.

- *φ and χ*: The pr function typically give a score between 0 and 0.1, with mean sitting at a much lower value than 0.1, and the standard deviation is less than 0.02. We would ideally have cost distribution with a higher standard deviation to differentiate link costs among different pairs of agents. Empirical evidence show that $\varphi(x) = x^{1/3}$ as an optimal setup. No method exists that quantifies the link cost between two disjoint networks. We set $\chi = 0.2$ to give link costs between agents in different networks a cost of 5.
- *α and λ*: Following the settings in [47], we set the parameter λ to 0.6 and the restart factor α to 0.9.
- *B*: We set B as a percentage of network size n to see how it influences performance. For synthetic networks, we vary B from $0.1n$ to $0.5n$ for robustness of our empirical findings. For DBLP networks, we vary B from $0.05n$ to $0.25n$.

Experiment setup. Each trial involves (a) deriving utility and cost by computing centrality and personalized PageRank, (b) picking up a random target v and (c) applying heuristics above to seek for a set S for v to build edges with.

We also perform a benchmark method and compare its performance with our heuristics for the U-NB. Prior to this work, link recommendation is usually done through link prediction. our goal is to test how link recommendation is affected once we take utility and costs into consideration. For the benchmark method, we pick a well-established link prediction algorithm, namely, the *resource allocation (RA) index*, which stands out as a superior similarity-based link predictor compared to many other alternatives [28]. The method is motivated by the resource allocation dynamics on complex networks. Assuming an agent x tries to send resources to y, who's not adjacent to x, the common neighbors of x and y play a role of transmitting the resource equally. The RA index of x and y is $s_{xy}^{\text{RA}} = \sum_{z \in N(x) \cap N(y)} 1/|N(z)|$.

The performance of a method is then evaluated based on utility of the resulting sets. To ensure robustness of results, we randomly select 50 agents as target v in each experiment and analyze average utility gained by the selected agents.

We first conduct experiments on networks generated by ER, BA and WS model with 100 nodes each. As the networks are connected, we use cost as in (3). Then for these models again, as well as for all DBLP networks, we apply the heuristics with the smoothed cost in (5). Comparison of results should illustrate consistency between the two different cost formulas.

Experiment 1 (Connected networks). As shown in Fig. 1, all three heuristics significantly outperforms the RA based benchmark. The plot shows utility U_G as a function of the budget B. The three heuristics SPP, SCLS, SUN are shown in green, yellow and red curves, respectively. Orange curves represent benchmark method (RA), and blue curves show average starting centrality. This color scheme will be used throughout all experiments. Clearly, all four methods give increased centrality, but RA performs poorest with an increase of around 1.34 times. SPP, SCLS and SUN achieve significantly better results. The difference between RA and other three heuristics increases as B increase. Note also that SUN and SCLS outperforms SPP by at least 10.3% and the difference among these three strategies increases as the budget grows, which indicates that SUN and SCLS have superior utility-maximizing ability than the link prediction-based method.

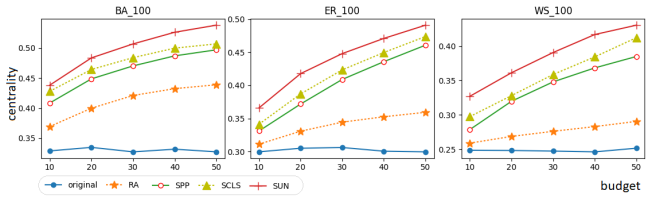

Figure 1: Performance in connected networks.

Experiment 2 (Smoothed cost). Fig. 2 shows the results for the ER, BA and WS networks over smoothed cost function. Results on networks are consistent with Exp. 1. For example, in network WS-500, setting $B = 0.5n$, RA improves average utility of the target from 0.118 to 0.216, while SPP, SCLS, SUN have utility 0.277, 0.294, and 0.368, resp. Our three heuristics improve the centrality of the target by at least 2.344 times, while RA only improves it by 1.824 times. Unsurprisingly, in Figure. 2 SUN and SCLS, which are based on the utility of a potential link and the closeness centrality of a node, outperforms SPP by at least 10.2% and at most 12.59%.

Experiment 3 (DBLP networks). Fig. 3 shows the results on DBLP networks utilizing SAN model. The results are consistent with the one on generated non-attributed networks. RA can increase the centrality of a target node for about 4.389 times, while SPP can enhance for about 5.336 times, SCLS 5.868 times and SUN 7.605 times with same budget, shown in the DBLP_$t2$ network when we set the budget to be 5% of the node number. In every budget, SPP, SCLS and SUN clearly outperform the benchmark method, and the improve by SUN is far more than SPP by at most 12.76%.

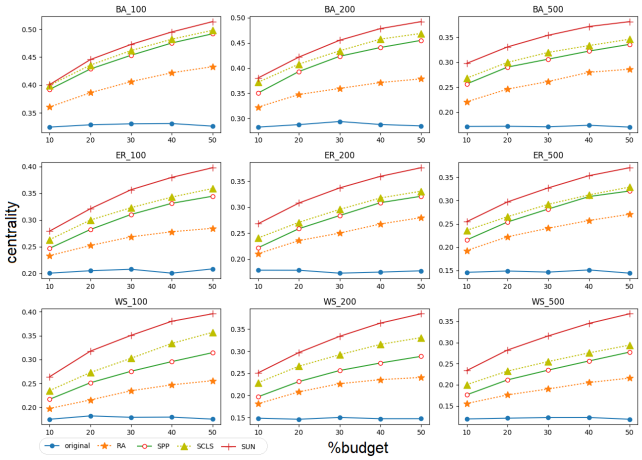

Figure 2: Performance in connected networks with smoothed costs

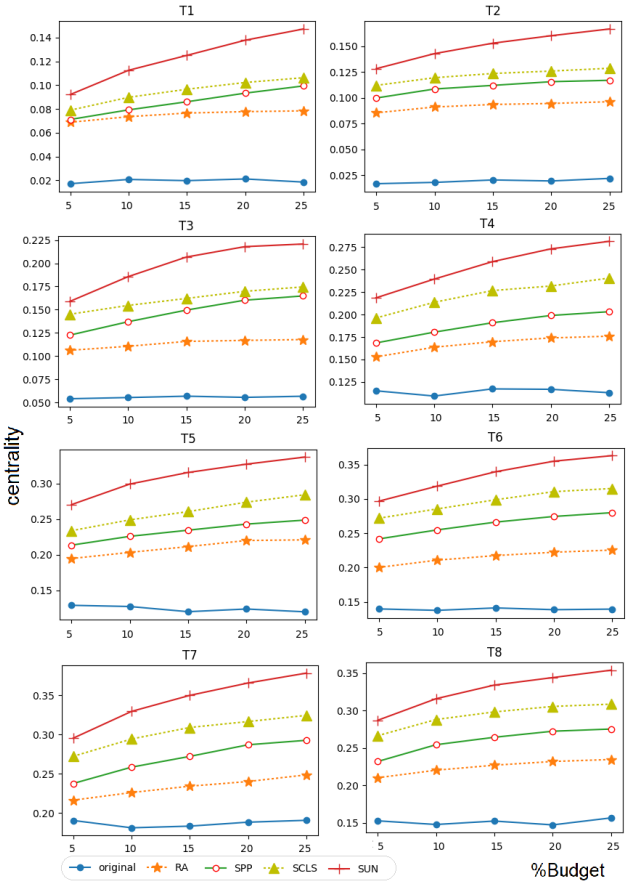

Figure 3: Performance in DBLP networks.

5 UTILITY-BASED NETWORK FORMATION MODELS

Our next goal is to propose network formation models that rely on network building strategies. The behavior of a social system is characterized by the collective actions of self-interested autonomous agents. In views of strategic network formation, links are established as each agent aims to maximize gain through building relations. A network structure takes shape as we allocate budgets amongst agents allowing them to establish links. To derive a model that is as lean and simple as possible, we assume: (1) No edge will be severed, so that focus is solely placed on tie creation. This is consistent with the classical BA model of scale-free networks; and (2) All agents are given the same budget B, and decide edges that they would form based on the current network structure.

Our network formation is defined as follows: Let $\mathcal{S}$ be a network building strategy for an agent. Given an initial network $G = (V, E)$, every agent $v \in V$ applies $\mathcal{S}$ and computes a set $S_v \subseteq V$ of nodes to establish links with. The resulting network is then

$$G' = (V, E \cup \{\{v, u\} \mid v \in V, u \in S_v\}) \tag{8}$$

Core/periphery structure. A core/periphery structure, informally speaking, is formed by a bipartition between a dense core and a comparatively sparse periphery, and that the graph cannot be decomposed further into several loosely connected dense clusters [9]. In this sense, a core/periphery structure can be viewed as being in the other end of a spectrum opposite to a community structure, which consists of several dense subgraphs but no core. The core must sit at the center of the network; This is a stronger property than scale-freeness, as e.g., a network with power-law degree distribution may easily contain a dense subgraph, however, the dense subgraph may not occupy a very central position relative to others and thus the network is not considered a core/periphery structure.

Here, we would like to test the existence of core/periphery structure in both G and the outcome G'. To this end, we recall an important definition: Given a network $G = (V, E)$, a *k-core* is a maximal subgraph with the minimum degree k [35]. The *core number* $\kappa(G)$ of G is the largest k such that G contains a k-core. For any subset $S \subseteq V$, use $\mathrm{cl}_G(S)$ to denote the closeness centrality of S, i.e.,

$$\mathrm{cl}_G(S) = \frac{|S| \cdot (n-1)}{\sum_{v \in S} \sum_{u \in V \setminus \{v\}} \mathrm{dist}_G(v, u)}. \tag{9}$$

Let $V_{\mathrm{core}}(G)$ denote the $\kappa(G)$-core with maximum closeness centrality. The following presents a Holme's well-accepted indicator of a core/periphery structure [16]:

Definition 5.1. Given $G = (V, E)$, $\mathcal{G}(G)$ denotes the collection of all graphs with the same degree sequence as G. The *core/periphery coefficient* of a graph $G = (V, E)$ is given by

$$c_{\mathrm{cp}}(G) = \frac{\mathrm{cl}_G(V_{\mathrm{core}}(G))}{\mathrm{cl}_G(V(G))} - \left\langle \frac{\mathrm{cl}_H(V_{\mathrm{core}}(H))}{\mathrm{cl}_H(V(H))} \right\rangle_{H \in \mathcal{G}(G)} \tag{10}$$

where $\langle \cdot \rangle$ denotes the averaging operation.

Intuitively, G has high core/periphery coefficient if its most central $\kappa(G)$-core has predominantly higher closeness centrality than the central $\kappa(H)$-core in a null model. If $c_{\mathrm{cp}}(G) > 0$, we can consider that G exhibit core/periphery structure, with the property becomes more evident as $c_{\mathrm{cp}}(G)$ increases. None of ER, BA and WS models

produce graphs with positive $c_{cp}(G)$. Several notable examples of networks with core/periphery structure include rich-club networks (whose $c_{cp}(G)$ is typically around 0.1) and onion networks (whose $c_{cp}(G)$ is around 0.2); See [9] for an introduction.

Experiment 4 (Regular graph). Naturally, the network G' depends on (a) the network building strategy $\mathcal{S}$ and (b) the initial network G. For $\mathcal{S}$, we may apply any of the strategies SCLS, SPP and SUN studied above. For G, we assume that upon initialization, the agents form a regular network where all agents have exactly 3 adjacent edges; this makes initial edges evenly distributed within G, and eliminates any prior distribution affecting the outcome. For this experiments, we generate regular graphs of size 1000 and run the models when $\mathcal{S} \in \{$ SCLS, SPP, SUN$\}$. Empirically we would like to investigate structural properties of the network G'. Particular focus will be put on the following desirable properties [48]:

- *Scale-freeness*: Power-law degree distribution $P(k) \sim k^{-\gamma}$ for degree $k \in \mathbb{N}$, with γ around 2 or 3.
- *Small-world property*: Small shortest path length (typically < 6), and High clustering coefficient (above 0.1).
- *Core/periphery structure*: A positive core/periphery coefficient (> 0.1).

See Fig. 4 for the resulting networks. Table 1 contains detailed indices for the initial network G as well as three resulting networks G':

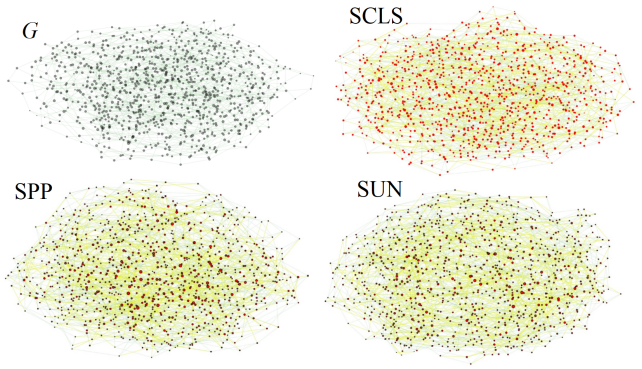

Figure 4: The result of the network formation model starting from a 3-regular graph and applying SCLS, SPP and SUN strategies. (top left) Initial 3-regular network; (top right) G' with SCLS; (bottom left) G' with *SPP*; (bottom right) G' with SUN.

The three resulting networks exhibit very similar patterns: Each network contains roughly 900-1000 new edges. The network all show clear power-law distribution (with $\gamma \in [2.6, 3.1]$), verifying that they shift from regular graph towards scale-free networks. This is reasonable as by adopting network building strategy, the agents engage in an essentially a more sophisticated form of "preferential attachment" that prefers more central nodes in the network. Among the strategies, SUN results in the highest γ indicating largest inequality among degree frequencies. There are significant increase on clustering coefficient and decrease in average path length, thus showing small-world property. The core/periphery coefficient remains very low, showing that none of the resulting networks have a core/periphery structure.

Table 1: Key indices for the initial regular graph G and the result G' of the network formation model with SCLS, SPP and SUN strategies. Goodness of fit indicates how well linear regression fits the log-log **degree distribution.**

	G	SCLS	SPP	SUN
Avg. deg	3.0	4.802	4.956	4.944
Avg. cluster. coef.	0.002	0.299	0.277	0.277
Avg. path. len.	8.077	5.981	5.870	5.880
Core/peri. coef.	-0.027	-0.019	-0.020	-0.020
γ	0	2.638	2.940	3.073
Goodness of fit	-	0.89	0.98	0.98

Experiment 5 (BA networks). We now augment our model by changing the initial graph G. In particular, we set G as a BA network produced by 1000 agents using preferential attachment. Similar to Experiment 4, we apply the three network building strategies to G. The resulting networks are shown in Fig. 5 with degree distributions displayed in Fig. 6. Key indices are listed in Table 2.

As shown in the results, each network contains roughly 700-900 new edges; this is smaller than the previous case where regular graph is chosen as more agents tend to choose the same edges to build due to unbalance of the initial network. All three resulting networks demonstrate clear power-law degree distribution and a hugely increased clustering coefficient (from 0.025 to over 0.5). Consistent with Experiment 4, neither of the SCLS and SUN strategies give rise to high core/periphery coefficient. For SPP, however, the core/periphery coefficient increases from -0.101 to 0.233, giving strong indication that a core/periphery structure emerges. An intuitive explanation to this phenomenon is this: as in SPP each agent creates ties that are predicted to be the easiest to establish for themselves, central nodes will tend to form links with other central nodes, while periphery nodes tend to connect to periphery nodes, amplifying the differences between these two levels.

Table 2: Key indices for the initial BA network G and the result G' of the network formation model with SCLS, SPP and SUN strategies. Goodness of fit indicates how well linear regression fits the log-log **degree distribution.**

	G	SCLS	SPP	SUN
Avg. deg	3.992	5.544	5.766	5.73
Avg. cluster. coef.	0.025	0.564	0.567	0.574
Avg. path. len.	4.113	3.514	3.555	3.396
Core/peri. coef.	-0.101	-0.056	**0.233**	-0.024
γ	1.454	1.206	1.224	1.104
Goodness of fit	0.845	0.799	0.793	0.763

6 CONCLUSION AND FUTURE WORK

We integrate costs and gains in the process of link creation. Gains to agents are measured in the form of improved closeness centrality;

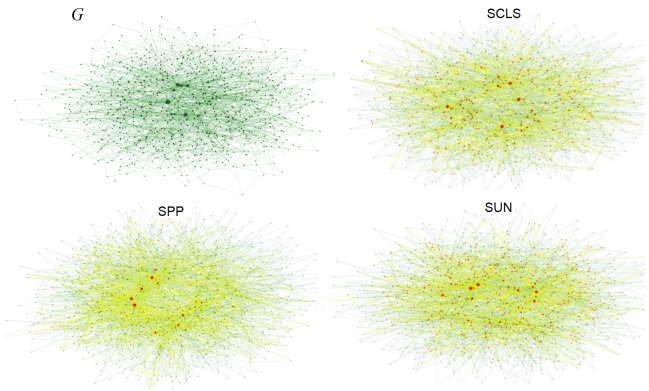

Figure 5: The result of the network formation model starting from a BA network and applying SCLS, SPP and SUN strategies. (top left) Initial network; (top right) G' with SCLS; (bottom left) G' with *SPP*; (bottom right) G' with SUN.

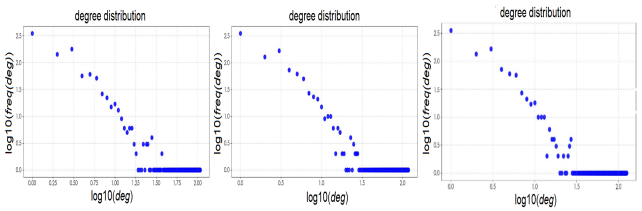

Figure 6: Degree distribution of G' with SCLS (left), SPP (center) and SUN (right) strategies.

costs of establishing ties are inversely correlated to link prediction scores. The two strategies that we propose, SCLS and SUN, result in superior performance compared to other strategies. In particular, exploiting submodularity of the utility function, SUN has a guaranteed approximation ratio of $\frac{1}{2} \times \left(1 - \frac{1}{e}\right)$. We then extend the problem from individual agents to simulating network formation over a set of utility-driven agents. We show:

(1) When starting from a regular graph, the agents would achieve both a power-law degree distribution and small-world property. Core/periphery structure, an important meso-level network property, is not present in this scenario.

(2) When starting from a scale-free network resulted from preferential attachment, the agents would also achieve the properties above. What's remarkable is that a core/periphery structure appears as the agents choose the cost-driven strategy SPP; such structure does not emerge for the other utility-driven strategies.

Ideas and methods proposed in this paper represent a novel research initiative. There are several potential directions for future work. A rather straightforward extension is to study utility-driven network building over dynamic networks. One may extend the current network formation model to a dynamic network model as follows: The process involves several rounds of evolution. With every round, the agents are assigned a certain amount of budget, and act by establishing links with others while keeping their costs within budget. Once a round is over, the agents are then assigned new budget, and start another round. This corresponds intuitively to an effect where agent may "reload" resources to create further links as time progresses. An interesting question to examine is to see whether core/periphery structure may emerge as the network evolves for multiple rounds.

A further extension is to also allow tie severance. Another potential future work is to choose alternative definitions of gains and costs. Rather than maximizing closeness centrality, an agent may have other desired outcomes such as increasing in influence. Apart from personalized PageRank, a different cost function may involve machine learning that computes easiness to create ties between individuals [22].

A long and fruitful research direction investigates network formation from a game-theoretical point of view. For example, Myerson's linking game assumes each agent as a rational player who is able to choose as strategies to link with others. Contrary to our framework, a link is only established if both players on its endpoints choose to build ties with each other. The advantage of this formulation is to allow the discussion of equilibria, i.e., stable states emerged as outcomes of the game. We speculate that our definition of utility may give rise to interesting network formation games.

REFERENCES

[1] Paul S Adler and Seok-Woo Kwon. 2002. Social capital: Prospects for a new concept. *Acad. Manage. Rev.* 27, 1 (2002), 17–40.
[2] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. *web search and data mining* (2011), 635–644.
[3] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.
[4] Francis Bloch and Matthew O Jackson. 2006. Definitions of equilibrium in network formation games. *International Journal of Game Theory* 34, 3 (2006), 305–318.
[5] Stephen P Borgatti and Martin G Everett. 2000. Models of core/periphery structures. *Soc Networks* 21, 4 (2000), 375–395.
[6] Gino Cattani and Simone Ferriani. 2008. A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the Hollywood film industry. *Organization Science* 19, 6 (2008), 824–844.
[7] James S Coleman. 1988. Social Capital in the Creation of Human Capital. *Amer. J. Sociology* 94, 1 (1988), 95–120.
[8] Pierluigi Crescenzi, Gianlorenzo Dangelo, Lorenzo Severini, and Yllka Velaj. 2015. Greedily Improving Our Own Centrality in A Network. *symposium on experimental and efficient algorithms* 9125 (2015), 43–55.
[9] Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi. 2013. Structure and dynamics of core/periphery networks. *Journal of Complex Networks* 1, 2 (2013), 93–123.
[10] Elise Desmier, Marc Plantevit, Céline Robardet, and Jean-François Boulicaut. 2012. Cohesive Co-evolution Patterns in Dynamic Attributed Graphs.. In *Discovery Science*, Vol. 7569. Springer, 110–124.
[11] Robin I M Dunbar. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* 16, 04 (1993), 681–694.
[12] Paul Erdös and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5, 1 (1960), 17–60.
[13] Martin Fink and Joachim Spoerhase. 2011. Maximum betweenness centrality: approximability and tractable cases. *workshop on algorithms and computation* (2011), 9–20.
[14] John Geanakoplos and Paul Milgrom. 1991. A theory of hierarchies based on limited managerial attention. *Journal of The Japanese and International Economies* 5, 3 (1991), 205–225.
[15] Neil Zhenqiang Gong, Ameet Talwalkar, Lester W Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Shi, and Dawn Song. 2014. Joint Link Prediction and Attribute Inference Using a Social-Attribute Network. *ACM Transactions on Intelligent Systems and Technology* 5, 2 (2014), 27.
[16] Petter Holme. 2005. Core-periphery organization of complex networks. *Physical Review E* 72, 4 (2005), 046111.
[17] Matthew O Jackson and Yves Zenou. 2014. Games on networks. (2014).

[18] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 137–146.

[19] Samir Khuller, Anna Moss, and Joseph Naor. 1999. The budgeted maximum coverage problem. *Inform. Process. Lett.* 70, 1 (1999), 39–45.

[20] Kate Larson. 2011. Social distance games. In *The International Conference on Autonomous Agents and Multiagent Systems*. 1281–1282.

[21] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne M Vanbriesen, and Natalie S Glance. 2007. Cost-effective outbreak detection in networks. (2007), 420–429.

[22] Xiaoyi Li, Nan Du, Hui Li, Kang Li, Jing Gao, and Aidong Zhang. 2014. A Deep Learning Approach to Link Prediction in Dynamic Networks. (2014), 289–297.

[23] Zhepeng Li, Xiao Fang, Xue Bai, and Olivia R Liu Sheng. 2017. Utility-Based Link Recommendation for Online Social Networks. *Management Science* 63, 6 (2017), 1938–1952.

[24] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[25] David Libennowell and Jon M Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology* 58, 7 (2007), 1019–1031.

[26] Nan Lin and Bonnie H Erickson. 2010. *Social capital: an international research program*. Oxford University Press.

[27] Jiamou Liu and Anastasia Moskvina. 2016. Hierarchies, ties and power in organizational networks: model and analysis. *Social Network Analysis and Mining* 6, 1 (2016), 106.

[28] Linyuan Lü and Tao Zhou. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications* 390, 6 (2011), 1150–1170.

[29] Feng Luo, Bo Li, Xiufeng Wan, and Richard H Scheuermann. 2009. Core and periphery structures in protein interaction networks. *BMC Bioinformatics* 10, 4 (2009), 1–11.

[30] H. W. Ma and A. P. Zeng. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 19, 11 (2003), 1423.

[31] Anastasia Moskvina and Jiamou Liu. 2016. How to build your network? a structural analysis. *Proc. of IJCAI 2016* (2016), 2597–2603.

[32] Anastasia Moskvina and Jiamou Liu. 2016. Integrating Networks of Equipotent Nodes. In *International Conference on Computational Social Networks*. 39–50.

[33] L Page. 1998. The PageRank citation ranking : Bringing order to the web. *Stanford Digital Libraries Working Paper* 9, 1 (1998), 1–14.

[34] Tuomas A Peltonen, Martin Scheicher, and Guillaume Vuillemey. 2014. The Network Structure of the CDS Market and its Determinants. *Journal of Financial Stability* 13 (2014), 118–133.

[35] Boris Pittel, Joel Spencer, and Nicholas Wormald. 1996. Sudden emergence of a giantk-core in a random graph. *Journal of Combinatorial Theory, Series B* 67, 1 (1996), 111–151.

[36] Puck Rombach, Mason A Porter, James H Fowler, and Peter J Mucha. 2017. Core-periphery structure in networks (revisited). *SIAM Rev.* 59, 3 (2017), 619–646.

[37] David A Smith and Douglas R White. 1992. Structure and dynamics of the global economy: network analysis of international trade 1965–1980. *Social forces* 70, 4 (1992), 857–893.

[38] Roshan Sumbaly, Jay Kreps, and Sam Shah. 2013. The big data ecosystem at Linkedin. In *Proceedings of the 2013 ACM SIGMOD*. ACM, 1125–1134.

[39] J David Symons, Ussama B Zaid, Christian N Athanassious, Adam E Mullick, Steven R Lentz, and John C Rutledge. 2006. Influence of Folate on Arterial Permeability and Stiffness in the Absence or Presence of Hyperhomocysteinemia. *Arteriosclerosis, Thrombosis, and Vascular Biology* 26, 4 (2006), 814–818.

[40] Hanghang Tong, Christos Faloutsos, and Jiayu Pan. 2006. Fast Random Walk with Restart and Its Applications. (2006), 613–622.

[41] Brian Uzzi and Shannon Dunlap. 2005. How to build your network. *Harvard Business Review* 83, 12 (2005), 53–60.

[42] Caroline S Wagner and Loet Leydesdorff. 2005. Network structure, self-organization, and the growth of international collaboration in science. *Research policy* 34, 10 (2005), 1608–1618.

[43] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *nature* 393, 6684 (1998), 440–442.

[44] Bo Yan, Yang Chen, and Jiamou Liu. 2017. Dynamic Relationship Building: Exploitation Versus Exploration on a Social Network. In *International Conference on Web Information Systems Engineering*. Springer, 75–90.

[45] Bo Yan, Yiping Liu, Jiamou Liu, and Yijin Cai. 2018. From the Periphery to the Center: Information Brokerage in an Evolving Network. *Proc. of IJCAI-ECAI 2018*.

[46] Erjia Yan and Ying Ding. 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the Association for Information Science and Technology* 60, 10 (2009), 2107–2118.

[47] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. 2010. A unified framework for link recommendation using random walks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, 152–159.

[48] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. 2014. *Social media mining: an introduction*. Cambridge University Press.