Causal Inference and Prediction for Network Data

Eric Kolaczyk (McGill University),
Elizaveta Levina (University of Michigan),
Tianxi Li (University of Minnesota),
Elizabeth Ogburn (Johns Hopkins University)

August 19, 2024 –August 23, 2024

1 Overview of the Field

The swift evolution of data collection technologies has yielded an abundance of network data across diverse fields such as social sciences, biology, neuroscience, and engineering. These networks represent complex systems where nodes (e.g., individuals, brain regions, genes) connect through edges (e.g., relationships, communication, or functional links), offering valuable insights into the underlying structures and dynamics of these systems [9, 11, 13, 24, 27, 30]. Yet, extracting meaningful insights from such data via probabilistic modeling and statistical inference remains challenging [1, 3, 7, 12, 15, 22, 23, 32], particularly inferring causal relationships [25, 10, 5, 4] and making predictions [17, 33, 29].

Network data poses unique challenges for prediction and causal inference due to dependence between connected nodes and potential confounding effects that violate traditional assumptions about training data and obscure causal connections [21, 8, 28, 6, 20]. For example, interactions within a social network or connectivity network may involve both direct influences and indirect effects mediated by other nodes. Despite significant advancements in addressing these challenges within predictive modeling [16, 19, 31] and causal inference [25, 2, 18, 14, 26], there remains a considerable gap between the need for adaptable, realistic analysis methods and currently available tools.

This workshop spotlighted how adapting random network models could help close these gaps by enabling more precise predictions that account for peer effects and supporting interpretable causal inferences. It aimed to bridge a critical divide at the intersection of causal inference, predictive modeling, and random network models. While statistical models such as stochastic block models, latent space models, and graphon models have proven effective in capturing network structure, their full potential for prediction and causal inference has not yet been realized. Historically, despite their intersecting applications in analyzing modern social systems, these research areas have advanced in relative isolation. This workshop sought to unite researchers across fields to investigate how statistical frameworks for network models could be extended or adapted for more flexible, robust, and practical applications in prediction and causal inference. Additionally, the discussions highlighted the need and generated ideas for innovative network models and inference tools, driven by real-world field experiments and empirical studies.

2 Workshop Format and Activities

The workshop was conducted in a hybrid format, with 31 in-person participants and 47 virtual attendees. The participants represented a diverse mix of experts from different backgrounds. Although the majority

were statisticians, the group also included economists and computer scientists, offering a broad range of perspectives on our central topics.

All talks were delivered by in-person participants, while online attendees actively engaged in discussions via Zoom and the workshop's Slack channel. The workshop featured the following key components:

- **Expository lectures:** Four extended lectures provided participants with a shared foundation for the workshop discussions. These lectures took place on the first day. Eric Kolaczyk and Keith Levin presented the first two lectures, introducing random network modeling and exploring contemporary applications in causal inference and prediction. The remaining two lectures, by Alexander Volfovsky and Dean Eckles, focused on causal issues in observational studies and network formulations, addressing a wide range of causal problems associated with network data.

- **Research talks:** Each in-person participant presented a research talk. These talks covered a range of topics, including novel research findings, real-world problem motivations, or reviews of specific subjects. Presenters had the autonomy to choose their content, and the sessions showcased some of the latest advances in relevant areas.

- **Slack channel:** With 78 participants attending in person or online, an efficient communication platform was critical for fostering meaningful interaction. A dedicated channel was created on *Slack* to facilitate discussions on both academic and networking topics. Participants could ask questions, engage in discussions, and share materials such as references and web links, organized by day and topic. The Slack channel substantially enhanced the level of interaction and served as a central hub for in-depth discussions.

- **Post-lunch discussion sessions:** After each lunch break, an hour was reserved for informal discussions. This time allowed participants to explore relevant topics and revisit questions raised during earlier sessions in a relaxed setting. These discussions fostered deeper exploration of ideas and connections that may lead to future collaborative projects.

- **Panel discussion:** The workshop concluded with an open panel discussion on the final morning. This session summarized the topics covered and addressed emerging open problems related to the workshops presentations. It also identified available data sources and discussed their implications for future research directions.

3 Presentation Highlights

3.1 Empirical Studies and Challenges in Causal Inference and Prediction in Networks

Dean Eckles presented research on the causal impact of local network structures, focusing specifically on "long ties" and their relationship to economic and social outcomes. His findings suggest that regions with a higher proportion of long tiesconnections without mutual neighborstend to have better economic indicators, such as increased median income. The research also highlights life events, such as interstate migration or attending multiple schools, as significant factors in the formation of these long ties. Experimental evidence indicates that network structures influence behaviors like cooperation and tie formation, with long ties promoting broader access to non-redundant information, facilitating economic resilience and growth. He situates this work within the broader discourse on social networks, referencing theories such as Granovetters strength of weak ties and Burt's concept of structural diversity as drivers of information and economic advantage. Despite these findings, open challenges remain, such as isolating individual-level causal effects from ecological correlations and understanding the role of long ties across varying network contexts. This research adds valuable insight into how network structure can influence both community outcomes and individual life trajectories.

In his talk, Sharmodeep Bhattacharyya presented an analysis of political polarization using survey and network data, focusing on the effects of the COVID-19 pandemic on voting patterns in the 2020 U.S. elections. His study finds that in counties with higher COVID-19 infection and death rates, voters tended to favor

incumbents from their own party (co-partisans) while opposing incumbents from the opposing party (contra-partisans), suggesting that the pandemic intensified partisan loyalty and opposition. Utilizing a statistical framework, the analysis incorporates pandemic-related metrics (peak infection and death rates) and controls for demographic, socioeconomic, and health indicators to assess their influence on vote share changes between the 2016 and 2020 elections. The model highlights the interaction between pandemic severity and pre-existing partisan alignment at the county level, with significant coefficients indicating how extreme events like a pandemic can reinforce affective and ideological polarization. Bhattacharyyas findings suggest that social networks and shared crises like COVID-19 can significantly alter voting behavior, underscoring the complex relationship between public health crises and political alignment, and pointing to potential interesting directions for future causal influence research.

Panos Toulis presented a large-scale field experiment on tax audit policies in inter-firm networks, aimed at understanding both direct and spillover effects of tax audits on business compliance. Conducted with the collaboration of the Inter-American Development Bank and a South American tax authority, this experiment involved randomizing audit notices to nearly the entire network of businesses in the country, encompassing around 500,000 firms. The experiments design, implemented in multiple waves with increasing intensity, utilized firm transaction data to establish network connections and leveraged a fully randomization-based approach to accommodate the network's complexity and heterogeneity. They employed Fisherian Randomization Tests (FRT) to test hypotheses on both direct and spillover effects, observing that while direct effects of audits on compliance were robust, spillover effects on neighboring firms were more modest but observable, particularly in firms with multiple audited trading partners. The results offer insights into optimizing tax policies by understanding peer effects in business networks. The field study and analysis also motivates new methods to be designed in refining randomization designs to maximize power for detecting subtle spillover effects across high-degree nodes.

3.2 Causal Inference in Networks

Keith Levin, in his expository talk, introduced the topics of peer effects and causal inference within networks, emphasizing the complexities network structures impose on classical regression models. His lecture delved into the dynamics of direct effects, contagion, and interference within network settings, with a specific focus on contexts such as disease spread, where outcomes are influenced by the behaviors of both individuals and their neighbors. He presented a potential outcomes framework to define and estimate treatment effects, underscoring the challenges of confounding and mediation within networks.

Arun Chandrasekhar presented his research on the sensitivity of diffusion estimates in networks to measurement errors, particularly illustrating how even minor inaccuracies in network data or initial conditions can lead to substantial errors in predicting diffusion outcomes. In particular, his study focused on impacts of the potential uncertainty of the source node in a diffusion process and the network structures. Rigorous characterization of such impacts are given in the theoretical results. His study examined the implications of these errors within models such as the COVID-19 SIERD framework and other diffusion processes, demonstrating that predictions can significantly deviate due to slight mismeasured links or uncertain initial conditions. This work highlights the importance of incorporating uncertainty within network models for intervention studies.

Elena Zheleva provided a high-level overview of her groups research on causal discovery in networks. This line of work treats network structures as relational data and particularly focuses on the challenges of inferring causal relationships from relational data that include cycles. Cycles in the causal problems in relation data tend to increase both data and learning complexity. Her work addressed the complexities inherent in relational causal models (RCMs) and introduced methods for learning these models, thereby enhancing accuracy in analyzing peer effects and heterogeneous treatments within social networks. The presentation included an in-depth review of a series of methods aimed at advancing causal inference for real-world applications, such as examining how social media interactions might influence behaviors like vaccine hesitancy.

Elizabeth Ogburns presentation centered on "nonsense associations", which lead to inaccurate estimation of associations. This problem is understudied in network analysis, but is a widely observed phenomenon in network data based on recent studies from her group, as network-induced dependence structures can create spurious associations between variables. To solve this problem, her research studies the impact of network topology on correlation matrices in the setting of Markov random fields, showing how similar network structures between two variables can cause overdispersion or biased estimates even in the absence of

direct confounders. This work highlights the potential risk of such nonsense associations in network settings, and provides new theoretical insights into these associations, opening avenues for assessing the reliability of causal inference in network-linked data.

Heejong Bong introduced a doubly robust non-parametric estimator for estimating causal effects under network interference, presenting the KECENI (Kernel Estimation of Causal Effect under Network Interference) method. This approach addresses the challenge of accurately estimating causal effects in networks where units not only experience direct treatment effects but also interference from neighboring unitsa scenario common in social networks and epidemiological studies. The KECENI method combines kernel smoothing with a doubly robust framework, ensuring that estimates remain unbiased and consistent even when some aspects of the model (the outcome or propensity score functions) are misspecified. This work offers a flexible tool that performs well even with limited treatment overlap across network neighborhoods, providing a more nuanced understanding of indirect effects in networks, potentially enhancing the design of interventions in areas like public health and online platforms. The remaining challenges in this estimation problem also suggest fertile ground for future research in improving robustness and scalability in network causal inference methods.

David Chois presentation focused on estimating the number of units in a network affected by the treatment status of others, addressing interference in settings where one unit's outcome depends on the treatments applied to others. Traditional causal inference methods often assume no interference, yet in many real-world networkssuch as in epidemics or social influence contextsinterference is unavoidable. Chois approach defines several estimands to characterize interference, including the number of units influenced by any treatment, by treatments of neighboring units, or by treatments applied at a greater network distance. Using inverse propensity weighted (IPW) estimators, he proposed methods to lower-bound these estimands, achieving consistent estimates without imposing restrictive assumptions about the network structure or interference patterns. His work introduces techniques to construct confidence intervals for these estimates, offering a practical solution to understand the scope of network interference effects without requiring specific assumptions about the network dependencies. Applications in social networks and public health illustrate how these methods provide insights into indirect effects, informing policy decisions where understanding both direct and indirect treatment impacts is critical.

Christopher Harshaw introduced a novel spectral experimental design approach to estimate direct causal effects in networked settings where interference from neighboring units complicates causal inference. The method specifically addresses scenarios where treatments administered to one unit may influence outcomes for nearby units, creating "exposure mapping" challenges that can obscure direct effect estimation. By leveraging the eigenvalues of the adjacency matrix, they optimally assign "desired" exposure levels to units while resolving conflicts through an importance ordering mechanism, which minimizes variance in exposure assignments. The design applies a modified HorvitzThompson estimator to enhance precision, even discarding some data where conflicts persist, ensuring robust estimation. This work provides finite-sample guarantees on estimator accuracy and develops a framework for confidence interval construction under interference conditions. Harshaws spectral approach holds promise for improving experimental designs in fields like public policy, economics, and digital A/B testing, where network interference is common and direct effect estimation is crucial.

Several other speakers, including Alexander Volfovsky, Ben Bloem-Reddy, and Daniel Sussman, also shared their recent work and insights within this theme.

3.3 Random Network Models

Ji Zhu presented a novel latent space model for hypergraphs, designed to capture complex polyadic relationships among nodes, extending beyond traditional dyadic interactions. This model, termed the Diversity and Popularity Hypergraph (DiPH) model, aims to address real-world scenarios where interactions frequently involve multiple entities, such as in social networks, co-authorship networks, or product co-purchases. Unlike conventional methods that simplify polyadic interactions into dyadic connections, the DiPH model preserves the full structure of these multi-node relationships. Each node is assigned a latent position to represent diversity in interactions and a popularity parameter to capture varying levels of node participation in hyperedges. By modeling hyperedges based on both node diversity and popularity, the DiPH model better mirrors natural settings where entities interact in groups rather than pairs. The model can serve as a flexible option to model

the relational information without reducing them into pairwise relations, which can potentially makes the analysis more realistic in real-world scenarios.

In her talk, Yinqiu He explored advancements in latent space models tailored for heterogeneous networks, emphasizing their utility in analyzing complex systems like transportation and social networks. Specifically, she presented models that capture both shared and unique structural features across layers in multifaceted networks. For example, in the New York Citi Bike dataset, her model analyzed temporal changes in connectivity patterns between stations, accounting for time-varying node activity while preserving a consistent latent structure. By introducing a framework for latent embeddings that can adapt to multilayered and time-dependent data, Hes approach enhances predictive accuracy and interpretability, especially for longitudinal and multiplex networks. Her model employs a semiparametric efficient estimation procedure, achieving robust parameter estimates even in highly heterogeneous data environments. The models success on datasets like Citi Bike and multi-faceted lawyer networks demonstrates its potential for diverse applications, including urban mobility and professional relationship dynamics.

Alessandro Rinaldo discussed the limitations of exchangeable network models, particularly focusing on Exponential Random Graph Models (ERGMs) and the challenges associated with their probabilistic consistency and degeneracy. Rinaldo highlighted that finite exchangeability, a fundamental property for network models, often fails to ensure consistency across networks of different sizes, leading to problematic behaviors such as model degeneracy and unpredictable asymptotic characteristics. In ERGMs, for instance, the probability of observing a network is defined in terms of invariant statistics (e.g., edge and triangle counts), yet the resulting models may exhibit extreme sensitivity to parameter choices, sometimes concentrating on trivial and degenerate graphs like the empty or complete graph. The result provides fundamental insights into the interaction between sparsity, exchangeability and the parametric random network models.

Eric Kolaczyk presented his research on branching factor estimation in noisy networks. It tackles the critical issue of measurement error in network data, particularly for applications in epidemic modeling. He demonstrated that noise in network edges significantly impacts the accuracy of branching factors and subgraph density estimates, which are essential for understanding spread dynamics in epidemics. By developing method-of-moments estimators, Kolaczyk introduces robust methods that correct for biases and variance induced by network noise, yielding reliable estimates under various network conditions, from sparse to dense and homogeneous to inhomogeneous structures. This approach enhances inference accuracy in epidemic models by enabling better handling of noisy data, which is especially relevant in scenarios where real-world networks often include inaccuracies. Kolaczyks framework opens new avenues for addressing noise-related challenges in network analysis, with ongoing challenges related to optimizing these estimations under dynamic network conditions and complex dependency structures.

Subhadeep Paul discussed his work on identifying peer influence while adjusting for latent homophily in networks. They introduce an Embedding Network Autoregressive Model (ENAR), which addresses the dual challenges of modeling networked time series data and estimating causal peer effects in the presence of latent homophily. Unlike traditional autoregressive models, ENAR incorporates latent variables to control for unobserved homophily, thus improving the accuracy of causal peer influence estimates and enhancing predictive performance in multivariate network-linked time series. His approach separates peer effects from latent factors that drive both individual outcomes and the selection of peers. This reduces bias in estimating peer influence and is applied to real-world data from therapeutic communities, where peer interactions may influence recovery outcomes. The model advances existing methods by allowing more flexible latent space dimensions, improving upon previous models like CNAR and NAR in both theoretical and practical settings.

Emma Zhang presented an innovative pseudo-likelihood method for fitting the Popularity-Adjusted Block Model (PABM), addressing the challenges of community detection in large-scale networks. The PABM is a very flexible model for networks with community structure and community-wise degree heterogeneity. However, the model fitting can be extremely difficult for large scale problems. The proposed approach decouples row and column labels in the adjacency matrix, allowing for separate modeling of nodes' popularity within communities. This is particularly beneficial in networks where node popularity varies widely across classes. The corresponding pseudo-likelihood function allows for iterative updates via an alternating maximization algorithm, efficiently optimizing parameters with closed-form solutions in each step. The method consistently achieves lower classification error and faster computation times compared to traditional methods, especially as network sizes scale up in experiments, showing notable improvements in community detection in real-world problems.

Several other speakers, including Can Le, Nynke Niezink, and Jess Arroyo, shared their recent insights and contributions within this thematic framework.

3.4 Statistical Tools for Network Inference and Prediction

Liza Levina presented her work on an interpretable network-assisted prediction method, designed to enhance prediction accuracy by incorporating network structure while retaining model interpretability. This approach addresses the challenge that many machine learning models assume data independence, which doesnt apply in networked data where sample dependencies are prevalent. Leveraging network cohesion, the model ensures that connected nodes exhibit similar behaviors, improving prediction accuracy without losing interpretability. The method balances the flexibility of complex models like random forests or deep learning with the simplicity of regression by enforcing cohesion through a penalty and/or adding latent node positions into the models. This framework allows to compute feature importance analyses at both global and local levels, providing transparency regarding which features or network connections drive predictions.

Carey Priebes presentation focused on spectral graph methods for detecting change points within dynamic network time series. He introduced the concept of a "Euclidean mirror, a low-dimensional representation of latent dynamics within network data that captures essential temporal changes. This Euclidean mirror is constructed by aligning spectral embeddings across time points and employing multidimensional scaling (MDS) and Isomap techniques to create a simplified trajectory of network structure over time. By analyzing these trajectories, the method can detect significant shifts or change-points in the network, such as those due to external disruptions or shifts in policy, as demonstrated through applications like pandemic-related changes in organizational networks. This approach provides a powerful tool for understanding how network structures evolve, as it enables the identification of transition points where network behaviors fundamentally change, offering potential applications in fields like social science, neuroscience, and finance. It also opens new paths to extend the approach to more complex, multilayer networks, where capturing and interpreting subtle temporal patterns and interactions presents ongoing challenges.

Vincent Lyzinskis presentation examined the impact of vertex label shuffling on network inference across multiple graphs, addressing the challenges posed by label mismatches in paired or multi-network data. Label shuffling, often due to errors in data sampling or preprocessing, can severely degrade the accuracy of tasks such as clustering, classification, and hypothesis testing within graph data. To mitigate these issues, they introduced methods leveraging seeded graph matching, where a subset of correctly aligned vertex pairs, or "seeds," is used to iteratively improve alignment accuracy. He discussed applications of these methods within frameworks like the RDPG model, emphasizing how seeded graph matching can enhance inference power by reducing ambiguity in vertex correspondence. However, the process remains complex and computationally demanding, with over-alignment and phantom alignments as potential pitfalls that can lead to misleading significance in statistical tests. This work highlights a critical open problem in network analysis: developing scalable and robust methods for handling label uncertainty, particularly in high-dimensional graph spaces and in the presence of structured noise.

Tracy Ke introduced a goodness-of-fit (GoF) test for network models based on cycle count statistics, employing a Self-Normalizing Cycle Count (SCC) approach to assess model adequacy across various network structures. This method calculates cycle counts for cycles of length $m \geq 3$, then normalizes these counts to yield a parameter-free test statistic that approximates a standard normal distribution under the null hypothesis, making it robust for model misspecifications. Ke demonstrated the SCCs effectiveness within the block model family, including stochastic block model (SBM), its degree-correct version and the mixed membership versions, where traditional models often struggle to represent networks with severe degree heterogeneity or mixed memberships. The SCC approach involves estimating an adjusted parameter matrix, under different null models, which allows for a direct comparison against observed network data. This work provides a robust framework for model validation in complex networks.

Shirshendu Chatterjee presented advanced methods for change-point detection in sparse dynamic network models, addressing both offline and online detection scenarios. His work focuses on estimating change points in sequences of networks, with applications to the general inhomogeneous Erdös-Rényi model, a widely used framework for modeling complex network data. Chatterjee introduced algorithms that leverage window-based methods and Wild Binary Segmentation (WBS) to locate change points, using adaptive and data-driven approaches to detect shifts in network structure. For offline detection, theoretical results demonstrate the con-

sistency of these estimators, along with optimality guarantees under specified signal strength conditions. In online settings, the methods detect change points sequentially as new data arrives, enhancing the practicality of this approach for real-time applications. These techniques are particularly useful in scenarios where edge probabilities vary significantly across nodes, offering robust tools for identifying structural changes in diverse applications such as social network analysis and monitoring evolving interaction patterns.

Several other speakers, including Karl Rohe, Yingying Fan, and Tianxi Li, also shared recent work and insights within this thematic framework.

4 Outcome of the Meeting

The primary outcome of the workshop was the productive exchange of ideas across disciplines, deepening participants' understanding of both statistical tools and open challenges. This cross-disciplinary interaction, highlighted as the most valuable aspect by attendees, was facilitated by the diverse backgrounds represented. For example, the workshop convened applied economists, engaged in large-scale field experiments with complex prediction and causal inference needs, alongside theoretical statisticians, who develop the foundational models for these analyses. These two groups rarely converge at the same venues, making this exchange uniquely beneficial. Participants gained insights into each others research fields and recognized the value of interdisciplinary perspectives.

The talks and panel discussions brought up several fundamental challenges and research directions. Based on the panel discussions and after-workshop feedback, below are some of the potential crucial problems for future research efforts.

1. **Heterogeneous Spillover Effect Estimation in Social Networks**: Social influence, which significantly impacts behaviors, trends, and norms, requires an understanding of spillover effects within networked settings. Traditional causal inference approaches struggle with interference in networks, where an individuals outcome may depend not only on their treatment but also on the treatments of their neighbors. More general and flexible methods are needed for these problems. For example, how to leverage graph neural networks (GNNs) with attention mechanisms to model complex, individualized spillover effects, addressing heterogeneity in social influence based on personal traits and local network structures, may be an important problem to study. By enhancing causal inference in the presence of network interference, research in this direction aims to provide robust methods applicable across fields such as epidemiology, political science, and economics.

2. **Inference of Diffusion Process in Attributed Social Networks**: The spread/diffusion of social impact plays a role in many settings, for example epidemiological modeling. Accurately identifying the diffusion mechanism and initialization region accurately is essential for effective intervention. This research direction focuses on developing flexible tools with theoretical accuracy guarantees and computational efficiency for diffusion inference in complex social networks. By integrating node-specific attributes, such as socioeconomic status and preferences, these methods can potentially improve inference robustness and adaptability.

3. **Modeling and Inference of Nonsense Associations**: Given the suspicious and misleading conclusions resulting from such associations, there is great demand for developing methods to quantify and mitigate variance inflation or deflation in association estimates caused by network dependence. Extending the current available results, new methods could be designed to diagnose and correct for these inflated associations under different network structures, enabling more accurate and conservative inference across a broader array of network contexts, such as social, biological, and spatial networks.

4. **Incorporating Randomness of Network Structures in Prediction and Causal Inference**: As highlighted in multiple talks during the workshop, errors in network data can introduce significant challenges to inference accuracy in both predictive modeling and causal analysis. Developing the next generation of statistical tools that account for the inherent randomness in network structures is crucial to addressing these issues. Such methods would require integrating both established and novel network models into the statistical inference framework. Given the additional dependence and complexity this

integration may introduce, new technical tools are likely needed, necessitating intensive collaboration across diverse research areas.

A few additional interesting problems were also discussed, such as preparing publicly available data sets for model evaluation, establishing collaborative platform for large-scale social effect experiments, and more flexible dynamic network models with categorical network-linked time series.

Additionally, the workshop inspired participants to refine their research agendas based on newly gained perspectives. Several attendees reported initiating collaborations with new colleagues they met at the event. This momentum suggests a forthcoming wave of research aligned with the workshop's central themes. A few junior researchers have even begun planning follow-up events to continue fostering communication and collaboration, ensuring the workshops impact endures.

5 Acknowledgments

The organizers extend their heartfelt thanks to the Banff International Research Station (BIRS) for sponsoring this workshop. We are deeply grateful for the timely and dedicated support from the BIRS staff, who created a warm and welcoming atmosphere that made all the difference for our participants. The friendly and inspiring environment fostered by BIRS was essential to the event's success, enabling truly meaningful and productive scholarly interactions for everyone involved.

Travel to BIRS for junior researchers was partially supported by NSF Focused Research Group grant 2052918, to Liza Levina, Keith Levin, Jianqing Fan, and Yingying Fan.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[2] Abdullah Almaatouq, Alejandro Noriega-Campero, Abdulrahman Alotaibi, PM Krafft, Mehdi Moussaid, and Alex Pentland. Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, 117(21):11379–11386, 2020.

[3] Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, and Yichen Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

[4] Abhijit Banerjee, Emily Breza, Arun G Chandrasekhar, Esther Duflo, Matthew O Jackson, and Cynthia Kinnan. Changes in social network structure in response to exposure to formal credit markets. *Review of Economic Studies*, 91(3):1331–1372, 2024.

[5] Abhijit Banerjee, Arun G Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O Jackson, Harini Kannan, Francine N Loza, Anirudh Sankar, Anna Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

[6] Guillaume W Basse and Edoardo M Airoldi. Limitations of design-based causal inference and a/b testing under arbitrary and network interference. *Sociological Methodology*, 48(1):136–151, 2018.

[7] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[8] Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55, 2009.

[9] Jennifer A Dunne, Richard J Williams, and Neo D Martinez. Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

[10] Kenneth A Frank and Ran Xu. Causal inference for social network analysis. *The Oxford handbook of social networks*, pages 288–310, 2020.

[11] Agata Fronczak and Piotr Fronczak. Statistical mechanics of the international trade network. *Physical Review E*, 85(5):056113, 2012.

[12] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

[13] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[14] Alex Hayes, Mark M Fredrickson, and Keith Levin. Estimating network-mediated causal effects via spectral embeddings. *arXiv preprint arXiv:2212.12041*, 2022.

[15] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

[16] Can M Le and Tianxi Li. Linear regression and its inference on noisy network-linked data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1851–1885, 2022.

[17] Tianxi Li, Elizaveta Levina, and Ji Zhu. Prediction models for network-linked data. *The Annals of Applied Statistics*, 13(1):132–164, 2019.

[18] Wenrui Li, Daniel L Sussman, and Eric D Kolaczyk. Causal inference under network interference with noise. *arXiv preprint arXiv:2105.04518*, 2021.

[19] Robert Lunde, Elizaveta Levina, and Ji Zhu. Conformal prediction for network-assisted regression. *arXiv preprint arXiv:2302.10095*, 2023.

[20] Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, pages 3700–3708. PMLR, 2021.

[21] Charles F Manski. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542, 1993.

[22] Soumendu Sundar Mukherjee, Purnamrita Sarkar, YX Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*, pages 10694–10704, 2018.

[23] Mark EJ Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315, 2016.

[24] MEJ Newman. Network structure from rich but noisy data. *Nature Physics*, page 1, 2018.

[25] Elizabeth L Ogburn, Ilya Shpitser, and Youjin Lee. Causal inference, social networks and chain graphs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(4):1659–1676, 2020.

[26] Elizabeth L Ogburn, Oleg Sofrygin, Ivan Diaz, and Mark J Van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

[27] Frank Schweitzer, Giorgio Fagiolo, Didier Sornette, Fernando Vega-Redondo, Alessandro Vespignani, and Douglas R White. Economic networks: The new challenges. *Science*, 325(5939):422–425, 2009.

[28] Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

[29] Christoph Stadtfeld, András Vörös, Timon Elmer, Zsófia Boda, and Isabel J Raabe. Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences*, 116(3):792–797, 2019.

[30] Xiao-Qian Sun, Hua-Wei Shen, and Xue-Qi Cheng. Trading network predicts stock price. *Scientific Reports*, 4(1):3711, 2014.

[31] Jianxiang Wang, Can M Le, and Tianxi Li. Perturbation-robust predictive modeling of social effects by network subspace generalized linear models. *arXiv preprint arXiv:2410.01163*, 2024.

[32] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

[33] X Zhu, R Pan, G Li, Y Liu, and H Wang. Network vector autoregression. *The Annals of Statistics*, 45(3):1096–1123, 2017.