

Essays on Network Analysis with Applications to Seeding and Art Valuation

Malek Ben Sliman

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2021

© 2021
Malek Ben Sliman
All Rights Reserved

ABSTRACT

Essays on Network Analysis with Applications to Seeding and Art Valuation

Malek Ben Sliman

The rise and growth of online social networks have spurred tremendous changes in our understanding of human behavior. Social scientists and companies have devised new tools to analyze the vast amounts of data obtained from these networks. Such advances have had two major consequences. First, it has allowed firms to significantly improve their segmentation and targeting strategies. Second, it also modified how problems are conceptualized. For example, books, academic papers, or webpages are now being studied under methods developed for social network analysis.

This dissertation contributes to both applications. Essays 1 and 2 describe efficient targeting strategies in situations where access to information or computing power is costly. Although existing “seeding” methods have been quite successful in social networks, they often do not account for firms’ limited computing power or assume that firms are omniscient. Essay 3 focuses on the art industry by conceptualizing paintings as items connected to each other in a network through their visual similarities. Indeed, we still do not perfectly understand what makes art financially valuable and even major auction houses are at awe when paintings are sold at prices multiple times higher than what they expected. In particular, we aim to quantify how an art piece’s visual features and historical importance may impact prices and assess how auction houses and their marketing efforts may modify how art is evaluated and valued.

This dissertation has three essays. In the first essay, we analyze how the

friendship paradox, which states that your friends have more friends than you, may be generalized to situations where relationships are asymmetric. Indeed, the result assumes symmetric relations: if two people are friends, then each is the other's friend. For social networks that satisfy this assumption (e.g., Facebook), the friendship paradox implies that firms can potentially achieve faster and more widespread diffusion of information by seeding it with the friends of a group of people than with people in the group itself. We generalize the result to allow one-sided (leader/follower) relations and examine the implications for seeding in social networks where messages can be sent only by a leader to his/her followers. We obtain necessary and sufficient conditions under which the highest number of followers is obtained by seeding with (1) leaders, (2) followers, and (3) individuals chosen by ignoring the distinction between leaders and followers. We examine the seeding implications of the results for a subset of Twitter users.

The second essay furthers our understanding of the friendship paradox and relates it to beta centrality and eigenvector centrality. We generalize the results to asymmetric relations, define two beta centrality measures and relate them to the singular vectors of the associated directed graph. Our first generalization shows that the expected number of k removed friends is no smaller than the expected number of $k - 1$ removed friends when k is an even number. Such a relation does not necessarily exist when k is an odd number. As k increases to infinity, the limiting value of the expected number of k removed friends converges to the largest eigenvalue of the associated undirected graph. We interpret beta centrality to be a weighted sum of an infinite series of the numbers of k removed friends. It approaches eigenvector centrality when the weighting parameter becomes arbitrarily close to the inverse of the limiting value of the expected number of k removed friends. We further generalize these results to asymmetric relations (say, between followers and leaders) that can be represented by directed graphs. We show that the

last person in a randomly selected alternating sequence of $2k + 1$ leaders and followers (followers and leaders) has no fewer followers (leaders) than the last person in a randomly selected alternating sequence of $2k$ followers and leaders (leaders and followers). As k increases to infinity, the expected number of leaders of the last person in a randomly selected sequence of $2k$ alternating leaders and followers converges to a value proportional to the largest singular value of the associated directed graph. Similarly, the expected number of followers of the last person in a randomly selected sequence of $2k$ alternating followers and leaders converges to a (different) value proportional to the largest singular value of the associated directed graph. We show that there is a reciprocal relation between the limiting expected values of leaders and followers. We generalize beta centrality to asymmetric relations and relate the limiting values of beta centrality scores for followers and leaders to the singular vectors of the associated directed graph.

The third essay focuses on the art market. Auction houses hold auctions regularly throughout the year. However, once or twice a year, art investors and wealthy consumers attend highly selective marquee events: day and evening sales. Those carefully designed and highly marketed events often generate a lot of excitement for connoisseurs as most paintings get sold for tremendous amounts of money. But what makes those paintings special? We investigate how art is evaluated across those three types of auctions. Specifically, we build a deep learning model to summarize the paintings into a low dimensional representation space where each factor encodes a specific feature of the paintings' aesthetics and further utilize those components to create "network" variables that will determine how influential and creative a painting is. We use those predictors in hedonic regression models to study how art returns differs across the three types of sales and subsequently analyze whether the paintings are evaluated differently. In particular, we find that paintings sold in evening sales generated an annualized return of 14.33% in the

period 1999-2018 – more than three times the returns of paintings sold in regular or day auctions. Finally, we adopt a propensity score matching approach to create a homogeneous population of paintings – based on their likelihood to be auctioned in an evening sale – to assess the causal impact of being featured in an evening sale and find that such highlight increases a painting’s price by almost \$6 million.

Contents

List of Tables	iv
List of Figures	vi
Introduction	1
1 Asymmetric relations and the friendship paradox	6
1.1 Introduction	7
1.2 Asymmetric friendship paradox	10
1.3 Implications for seeding	13
1.4 Seeding on Twitter	20
1.5 Conclusion	26
2 Friendship Paradox Generalizations and Centrality Measures	28
2.1 Introduction	29
2.2 Symmetric Relations	32
2.3 Asymmetric Relations	37
2.4 Conclusion	46
3 The Impact of Auction Houses on Art Valuation	48
3.1 Introduction	49

3.2	Literature Review	52
3.2.1	Art as an Investment	52
3.2.2	Computer Vision	54
3.2.3	Social Network Analysis	55
3.3	The Dataset	56
3.3.1	Dependent Variables	58
3.3.2	Auction-level Variables	58
3.3.3	Painting-level Variables	59
3.3.4	Visual Features	59
3.3.5	Network Variables	66
3.4	Model of Art Valuation	71
3.4.1	Hedonic Regressions	71
3.4.2	Propensity Score Matching	79
3.5	Conclusion and Future Research	82
	Conclusion	84
	Bibliography	86
	Appendix A: Essay 1	92
	A1: Proof of Theorem 1	92
	A2: Calculations of $\mu_{f/l}$, $\mu_{f/f}$ and $\mu_{f/fr}$ for the graphs in Figure 1.2	95
	Appendix B: Essay 2	97
	B1: Proof of Theorem 1	97
	B2: Proof of Theorem 2	98
	Appendix C: Essay 3	101
	C1: VAE Architecture	101
	C2: Full Regression Results	102

C2a: Regular Auctions	102
C2b: Day Auctions	107
C2c: Evening Auctions	112
C3: Matching Performance	117

List of Tables

1.1	Average number of followers per seed obtained by seeding with leaders, seeding with followers, and undirected seeding for the relations shown in Figure 1.1	16
1.2	Average number of followers for different seeding methods using the largest network component and the screened subset of users	22
3.1	Visual factors and representative paintings	64
3.2	Cross tabulation between influence predicted by <i>theartstory.org</i> and G^{other} . Elements in the brackets correspond, respectively, to expected value and cell χ^2	69
3.3	Model performances for regular, day and evening auctions	76
3.4	Full model results for regular, day, and evening auctions	78
1	Panel (b): $\mu_{f/l} > \mu_{f/fr} > \mu_{f/f}$	95
2	Panel (c): $\mu_{f/fr} > \mu_{f/l} > \mu_{f/f}$	95
3	Panel (d): $\mu_{f/f} > \mu_{f/l} > \mu_{f/fr}$	96
4	Panel (e): $\mu_{f/f} > \mu_{f/fr} > \mu_{f/l}$	96
5	Panel (f): $\mu_{f/fr} > \mu_{f/f} > \mu_{f/l}$	96
6	Regression Results for Regular Auctions	102
7	Regression Results for Day Auctions	107

8	Regression Results for Evening Auctions	112
9	Mean Comparisons between Control and Treatment Group after Matching	117

List of Figures

1.1	Example of a directed graph G (left) and the associated undirected graph H (right)	15
1.2	Directed graphs illustrating the six possible orderings of the average number of followers per seed	18
1.3	Relation between number of followers and average number of tweets per day	20
1.4	Sample distributions of the total number of followers for 1,000 random users, random friends, random leaders, and random followers. Each sample distribution is based on 10,000 random draws.	23
1.5	Average number of followers (F) and average reach (R) for random users, random followers, and random leaders as a function of the size of the seeding sample (s) across 10,000 runs	25
1.6	Average duplication (D) for randomly selected users, leaders, followers, and friends as a function of the size of the seeding sample (s) across 10,000 runs	25
2.1	Networks for which $E(d(X_1)) > E(d(X_2))$ (left) and $E(d(X_1)) < E(d(X_2))$ (right)	34
2.2	Example illustrating forward and backward walks in a directed graph	39

3.1	Distribution of $\ln(\text{Price})$ for day, evening, and regular auctions	72
3.2	Evolution of the hedonic price indices for day, evening, and regular auctions	75
1	Variational Autoencoder Architecture. Transposed convolution layers have a kernel with 64 filters and a stride of size 2; remaining parameters: (kernel size, padding).	101

Acknowledgments

I want to express my gratitude to the many individuals who shaped my academic and personal path thus far. We tend to forget how luck and fate impact one's success. In my case, it was through those formidable people. I want to start by thanking my professors and mentors. My first thought goes to Professor Turbergue who inspired me to pursue Mathematics and "classes préparatoires." His support was instrumental to the successes I later encountered and I still heed his advice. Second, I want to thank Professors Jouini, Laux and Mayo Quenette from Ecole Centrale Paris. They were an important support to me both at the beginning and throughout my time at Columbia. Finally, I want to thank the Marketing department at Columbia. In particular, Professor Schmitt for his guidance, Professor Ascarza for her insightful comments (and her great sense of humor) and Professor Netzer with whom I enjoyed conversations about research and soccer alike. Last but not least, I want to thank my committee members starting with my two advisors. To Professor Jedidi, thank you for your trust and for "converting" me to Marketing. You have been a constant resource and I am extremely grateful. To Professor Kohli, I can't thank you enough for your openness. Thank you for helping me shape my research identity and for always offering to help when needed. I believe we share a similar vision of research and I am looking forward to our next

projects! To Professor Ansari, who chaired my committee, thank you for your humor and caring so much about Ph.D. students. To Professors Toubia and Iyengar, thank you for your great insight and feedback on my work.

I also want to thank all the friends who accompanied me over the years. Thanks to Sarra, William and Aziz with whom I shared so much during middle school and high school in Tunisia. Thanks to Selim, Shaheen, Aicha and Fares for the great moments we spent in Tunisia and Paris. We played pool, chess and trivia, went fishing, and traveled together and I am sure we will have as much fun in the future. Thanks to Ed who initiated me to beer, whiskey, good food, cricket and so much more at Centrale. Thanks to Nathaniel, Steven, Samuel and, last but not least, Mathias. Aside from all the fun times we shared, your ambition and confidence were extremely contagious. Thanks to Tarik, Emmanuel and Victor for initiating me to cinema and for all the great sports conversations. I also want to thank the numerous people that contributed to making New York a home. Thanks to Jenny who aside from all her help introduced me to Carmen and Gerald for his great support. Thanks to Richa, Charles, Kelthoum, Camille, Benoit, Gaetan and the many others I met during my master's degree. From Uris library to Miami, what a year! I also want to thank my roommates Lam-Ha, Magaly, Rupayan, Matt, Paula with a special thanks to Andy who inspired some of my work in this thesis and Daniela with whom Carmen and I spent a great time in quarantine between the Shore and the Sopranos. I have lived a true New York experience with you. Finally, I want to thank the people I met during my Ph.D. starting with Amanda, Vee, Chung, Luisa, Luz, Dan, Elizabeth, Verena, Khaled, Yanyan and Jasmine for listening to my jokes and helping me over the years. Thanks to Alain whose honesty and commitment to good research were always inspirational. Thanks to Nicolas for being so thoughtful in your feedback and for spending so much time talking about football. Having post-seminars (and a much needed cup of coffee) with the two of

you was quite an experience and I learned so much from those! Of course, thanks to Ma'ayan! In addition to deciphering the “game”, you always have a great spirit and attitude and I am looking forward to seeing all your successes.

I want to conclude this section by giving a special word of thanks to two amazing friends. To Messire Stéphane, I will always remember when we first met at Columbia and instantly became friends. From that moment on, we shared Ph.D. classes, started a radio show, watched major football games together, initiated countless jokes and micro-trends. You have been there for me throughout my entire Ph.D. journey, listening to my research ideas and inspiring confidence when I needed it most. We both know the importance of the 12th player and our “causeries” were truly legendary. To Valentin, so much must be said! Thank you for everything! You are by far one of (if not the) the smartest (how can't I mention how you reinvented one of Sylow's theorems from scratch in a few hours?) but also one of the nicest and most generous person I have ever met. We partied and worked hard together over the years and it was always a great time! During my Ph.D., we traveled quite a bit (notably Kentucky) and, most importantly, you were a tremendous support as you had to endure most of my complaints (and there were many). I am extremely lucky for your friendship and I will never thank you enough. I am hopeful that we do end up working and building a great project together.

As I transition to this last section dedicated to my family, I want to thank Bilel. We met over 20 years ago and since then so much happened but our friendship remained strong across Tunisia, Paris and New York. You made me discover the world under a different angle as we roamed through Tunis and as I discovered your entrepreneurial spirit. I grew tremendously from all those adventures. Over the years, you (and your family!) have always supported me and were always there for me. I still remember vividly all the nights we spent with a

couscous and some boukha which offered much needed distraction. You are not only a great friend but also family to me and I want to thank you for all of this! I am looking forward to celebrating with you!

Finally, I want to thank my family (and the numerous pets we have had!). My first thoughts go to my ancestors who paved the way to my success as well as my brother's. Among many other teachings, their stories taught me to have ambition, to dream big and to study and work hard; but also the importance of taking risks and being in control of your own life decisions. For instance, it would be hard not to mention my great-grandfather who traveled from his native Djerba to so many other places or how my grandfather became a successful lawyer after forty. My grandmother in particular had a very unique impact on me and not a day goes by without thinking of her as I carry her history and her words. I can only hope that she would be proud of me. Second, I want to give a special thanks to my Uncle Noé for his amazing support. Third, I want to thank my parents who were an example of tolerance and hard work and my brother. My father Jalel who has always been a role model for me (and who, in my eyes, is the funniest person on Earth). I could not be prouder of being a Ben Sliman! My mother Sarah who has always been a constant support and taught me how to work well. It is impossible to list all she did to contribute to our successes. It is, in fact, not a surprise that I ended up pursuing a Ph.D. in this field and she deserves the credit for it. My brother Sélim who used to quiz me on my lessons and who always inspired me for his wit and innate intelligence. He is a born leader and I aspire to be as calm as he is! My final thoughts are dedicated to Carmen who has been by my side from the start of my Ph.D throughout both good and exciting moments and periods of doubt. Although her name will not be on the diploma, I know how long and full of sacrifices this journey was. To Carmen, thank you for being my first supporter over the years, for accepting so many of my quirks, and, most importantly, for thinking I am funny!

Dedicated to my family and friends

Introduction

Since Myspace reached one million monthly users in the mid-2000s, use of social media has skyrocketed with multiple websites attracting several hundred million active users and Facebook and YouTube reaching more than 1.5 billion users every month. Those companies have entirely modified individuals' access to data while firms better cater products and content to customers reaching them through much simpler channels. In addition, by designing new network analysis methods to analyze the resulting datasets, scholars can now study how items are connected to each other. Similarly, the three essays in this thesis aim to use those network analysis approaches to analyze how people and items are connected to each other. The first two essays focus on the friendship paradox, which states that, on average, your friends have more friends than you, in the context of seeding and centrality. The friendship paradox provides an efficient method to target customers in situations in which access to data about the customers is limited. The third essay focuses on fine art auctions and conceptualizes art pieces as items connected to each other based on visual similarities, which allows us to assess how creative and influential individual paintings are. Using this modeling approach, we aim to understand how auction houses influence how art is valued through their marketing

efforts.

In the first essay, we analyze how the friendship paradox generalizes to unreciprocated relationships such as love, hate, and friendship. For social networks in which relationships are symmetric (e.g., Facebook), the friendship paradox implies that firms can spread information faster and to more people by adopting a two-step approach: first, randomly selecting people in the population and then sharing information with randomly selected friends of those persons. We generalize the results to allow one-sided leader/follower relations and examine how seeding (that is, generating a spreading phenomenon by only sharing information with a select few users) in those networks where information can flow only from a leader to followers can be achieved. We obtain necessary and sufficient conditions under which the greatest number of followers is obtained when seeding with (1) leaders, (2) followers, and (3) individuals chosen by ignoring the distinction between leaders and followers. We further simulate the theoretical results using a subset of Twitter users.

In the second essay, we deepen our understanding of the friendship paradox and show that it is related to the notions of beta centrality and eigenvector centrality which are measures of influence in a social network. We generalize these centralities to asymmetric relations by defining two beta centrality measures and relating them to the singular vectors of the associated directed graph. First, we show that the expected number of k removed friends is no smaller than the expected number of $k - 1$ removed friends when k is an even number and that this relation does not necessarily hold when k is an odd number. As k increases to infinity, the limiting value of the expected number of k removed friends converges with the largest eigenvalue of the associated undirected graph. Thus, we interpret beta centrality as a weighted sum of an infinite series of the number of k removed friends. The results are extended to asymmetric relations that can be represented by directed graphs. In

that context, we show that the last person in a randomly selected alternating sequence of $2k + 1$ leaders and followers (followers and leaders) has no fewer followers (leaders) than the last person in a randomly selected alternating sequence of $2k$ followers and leaders (leaders and followers). As k increases to infinity, the expected number of leaders of the last person in a randomly selected sequence of $2k$ alternating leaders and followers converges to a value proportional to the largest singular value of the associated directed graph. Similarly, the expected number of followers of the last person in a randomly selected sequence of $2k$ alternating followers and leaders converges to a (different) value that is proportional to the largest singular value of the associated directed graph. We show that there is a reciprocal relation between the limiting expected values of the numbers of leaders and followers.

In the third essay, we focus on the art market and aim to understand how the type of event at which a painting is auctioned influences its monetary valuation. Since 2009, the art market has grown by 62% in value and 34% in volume and, as of 2019, it had an estimated total value of around \$64.1 billion. However, our understanding of valuations in this industry is incomplete and art specialists are sometimes surprised by the amounts obtained by certain paintings. Many factors contribute to difficulty in predicting art prices. First, paintings are, by definition, unique and are rarely sold so extracting meaningful information from the paintings is key. Second, art buyers are extraordinarily heterogeneous, and their behaviors are driven by both investment values and aesthetics. Finally, the industry is notoriously unregulated, limiting access to data.

In particular, the extent and impact of marketing efforts by auction houses to advertise art pieces are not observed. In this essay, we focus on the type of auction at which a painting is sold. Some auction houses, for example, organize rare

marquee events that showcase a carefully chosen set of paintings. By focusing on differences in auctions, we contribute to understanding the degree to which auction houses directly influence sales prices of paintings.

Our analysis is based on a rich dataset containing observations for every fine art auction conducted in New York City from 1999 through 2018. The dataset covers more than 140,000 auctions from about 19,500 artists and across 11 auction houses. It contains information on the artists (e.g., country, birth/death), the paintings (e.g., medium used), and the auction (e.g., realized price) and images of the paintings. To incorporate the aesthetics of a painting in our model, we developed a variational autoencoder to summarize the paintings in a low-dimension representation space in which each factor encodes a specific feature of the paintings. We use these components as descriptors of the paintings and to predict which features make paintings more or less pricey. The components are also used to create a network of paintings that allow us to determine how influential and creative a painting is relative to other paintings.

We conceptualize the paintings' characteristics by assuming that creative paintings are both novel (dissimilar to pieces painted before them) and influential (similar to later pieces). We build a weighted directed network of paintings in which the weights correspond to the similarity (computed using the latent factors describing the paintings' features) of two paintings and the direction is temporal. Therefore, a painting from 1800 will point toward a painting from 2000. Hence, the weighted out-degree, which measures a work's total similarity to more recent paintings (its influence), should have a positive impact on the auction price while the weighted in-degree, which measures the work's total similarity to older paintings (a measure of its novelty) should have a negative impact on the auction price. In other words, an influential painting has a large weighted out-degree since many subsequent

paintings are similar to it and a creative painting has a small weighted in-degree because it is significantly different from previous paintings.

The variables are used in hedonic regression models to study how art returns from three types of sales (regular, day, and evening auctions) differ and, subsequently to analyze whether the paintings were evaluated differently. In particular, we find that paintings sold at evening auctions generated an average annualized return of 14.33% in the study period (1999-2018) – more than three times the returns on paintings sold in regular and day auctions. We then specifically analyze the impact of an auction house’s decision to feature a painting in an evening sale instead of a day sale. We adopt a propensity score matching approach to identify homogeneous populations of paintings and find that highlighting a painting in an evening auction increases the price obtained in the auction by an average of almost \$6 million.

Asymmetric relations and the friendship paradox

This paper is jointly authored with Rajeev Kohli.

1.1 Introduction

The friendship paradox states that, on average, your friends have more friends than you (Feld, 1991). It implies that things like diseases and information that spread by contact between people can do so faster among the friends of a group of people than within the group itself. For example, Christakis and Fowler (2010) found that friends of a group of students were more susceptible to catching the flu and caught it sooner than the students. Garcia-Herranz et al. (2014) showed that tracking infections among the friends of a randomly chosen group of individuals can provide an early warning of an epidemic outbreak. In marketing, Singer (2016), Kumar et al. (2018), and Kumar and Sudhir (2019) observed that the friendship paradox has implications for how a firm can initiate viral marketing campaigns: instead of seeding information about a product, promotion, or event with a random number of individuals in a target population, seed it with their better connected friends.

Unlike seeding methods based on network centrality (Tucker, 2008; Goldenberg et al., 2009; Libai et al., 2013) and opinion leadership (Iyengar et al., 2011), seeding based on the friendship paradox can be useful when firms do not have access to information about targeted individuals' ties and when such information is difficult and/or costly to obtain. Thus, it is useful for social networks like Facebook and Twitter that have millions of users and constantly changing relationships and memberships. The computational time and cost of identifying current opinion leaders and individuals with strong network centrality can be large, especially if the networks must analyze the data separately to match new profiles of target users for each client.

The friendship paradox assumes symmetric relations. When two people are friends, each is the other's friend. We extend the paradox to situations in which

people have asymmetric relations. In the context of viral marketing, the relevant distinction between symmetrical and asymmetrical relations is how information is sent. Two people in a symmetric relation can send information to each other. In an asymmetric relation, only one type of person can send information to the other. Without loss of generality, the sender is called the “leader” and the receiver is called the “follower.”¹ We define a symmetric network as one in which all (pairwise) relations are symmetric and an asymmetric network as one in which *at least one* relation is asymmetric. For example, Facebook is a symmetric network because a user can send information to and receive information from all friends. Twitter is an asymmetric network because users can send information only to followers and receive information only from leaders (people they follow). From a seeding perspective, persons who haul more followers in an asymmetric network and more friends in a symmetric network can send viral information to a large number of people. We examine the following questions regarding asymmetric networks:

1. On average, do your leaders have more followers than you?
2. On average, do your followers have more followers than you?
3. On average, do people with whom you have ties (that is, who are your leaders and/or followers) have more followers than you?
4. What are the conditions under which people in each of the preceding three groups — those with whom you have ties, your followers, and your leaders — obtain the greatest number of followers?

¹Though we emphasize asymmetry of information flow because of our interest in viral marketing, our results hold more generally for any type of asymmetric relation. For example, they hold when one person considers another to be a friend and the other does not (Carley and Krackhardt, 1996; Ball and Newman, 2013) and for other types of relations such as knowing, influencing, loving, liking, and disliking others and “hierarchical” relations such as those between parents and children, and employers and employees. The “leader” and “follower” labels can be arbitrarily (but consistently) assigned to the two types of people in an asymmetric relation. For example, we could call a parent a leader and the child a follower or reverse the labels. All our results are symmetric in the sense that they hold when the labels are exchanged.

We use the answers to these questions to evaluate the effectiveness of seeding with leaders, seeding with followers and undirected seeding, that ignores the distinction between leaders and followers when selecting a seeding sample.

Following is a summary of the results.

1. On average, your leaders always obtain more followers than you. Thus, seeding with leaders always obtains a greater average number of followers per seed than random seeding (seeding with a random group of individuals in a target population).
2. On average, your followers obtain more followers than you if and only if the covariance of the number of leaders and followers is positive — that is, if people with more leaders also tend to have more followers. In that case, seeding with followers obtains a greater average number of followers per seed than random seeding.
3. On average, people with whom you have ties obtain more followers than you if and only if the covariance of the ties and the number of followers is positive (people with more ties also tend to have more followers). In that case, undirected seeding obtains a greater average number of followers per seed than random seeding.
4. On average, your leaders obtain more followers than you, your followers, and people with whom you have ties (people who are your leaders and/or followers) if and only if the variance in the number of followers exceeds both the covariance of the number of leaders and followers and the covariance of the number of followers and number of friends, multiplied by a scaling constant. In that case, seeding with leaders obtains a greater number of followers per seed than random seeding, seeding with followers, and undirected seeding. We obtain similar conditions for the other seeding methods.

We empirically evaluate the seeding methods using a subset of Twitter data and find that seeding with leaders, seeding with followers, and undirected seeding obtain substantially more followers per seed than random seeding. For example, seeding with 50 leaders obtains an average of more than 80,000 followers, which is almost twice the average number of followers obtained by seeding with as many as 1,000 random individuals. We also examine how screening criteria for seeding affects the method that can obtain the greatest number of followers per seed.

We also characterize the empirical distributions of the number of followers obtained using each type of seeding methods and find that samples for which the number of followers are in the right tails of the distributions can be obtained by choosing the best seeding sample from multiple draws. We decompose the total number of followers obtained by each seeding method into the product of reach and duplication, where reach refers to the number of unique followers to whom a seeding sample can send information, and duplication refers to the average number of seeded individual from whom a follower can receive the information.

In the next section, we generalize the friendship paradox to asymmetric relations. We then discuss implications of the results for seeding in asymmetric networks. Finally, we present our analysis of Twitter data.

1.2 Asymmetric friendship paradox

We represent a network with asymmetric relations using a directed graph $G(N, A)$, in which N is the set of nodes and $A = \{(i, j) | i, j \in N\}$ is the set of arcs. An arc (i, j) from node i to node j means that person i is a follower of person j ; equivalently, person j a leader of person i . The graph G represents symmetric

relations when an arc from any node i to another node j is always accompanied by an arc from node j to node i . In that case, all results reduce to the results obtained for symmetric relations.

Let $n = |N|$ denote the number of individuals and $m = |A|$ the number of leader-follower relations. Then, $\mu_l = \mu_f = m/n$ where μ_l denotes the average numbers of leaders per person and μ_f denotes the average number of followers per person.

Let x denote a variable of interest, such as the number of followers or leaders of an individual. It is not necessary the variable relate directly to the network. For example, the variable could be an individual's wealth or the number of messages the individual sends to others. Suppose we associate a value of x with each vertex (individual) in a symmetric network. Let μ_x denote the average value of x across all individuals, μ_{fr} the average number of friends per individual, σ_x^2 the variance in x , and $\rho_{x,fr}$ the correlation between x and the number of friends. Eom and Jo (2014) generalized the friendship paradox and showed that

$$\mu_{x/fr} = \mu_x + \frac{\rho_{x,fr}\sigma_x\sigma_{fr}}{\mu_{fr}} = \mu_x + \frac{\sigma_{x,fr}}{\mu_{fr}} \quad (1.1)$$

where $\sigma_{x,fr} = \rho_{x,fr}\sigma_x\sigma_{fr}$ is the covariance of x and the number of friends. We further generalize this result to a directed graph, G , and use it to examine the effects of seeding with leaders and followers.

Extending the preceding notation, let $\mu_{x/l}$ ($\mu_{x/f}$) denote the average value of x per leader (follower) and $\sigma_{x,l}$ ($\sigma_{x,f}$) the covariance of x and the number of followers (leaders). Theorem 1 characterizes the relations between the average value of x across all individuals, the average value per leader, and the average value per follower. The Appendix provides a proof of the theorem.

Theorem 1. *The average value of x per leader is given by*

$$\mu_{x/l} = \mu_x + \frac{\sigma_{x,f}}{\mu_f} \quad (1.2)$$

and the average value of x per follower is given by

$$\mu_{x/f} = \mu_x + \frac{\sigma_{x,l}}{\mu_l}. \quad (1.3)$$

Theorem 1 reduces to Eom and Jo's (2014) result in the special case in which G represents a symmetric network (that is, when each arc (i, j) is accompanied by the arc (j, i)). Note that $\mu_{x/l} = \mu_x$ when $\sigma_{x,f} = 0$ and $\mu_{x/f} = \mu_x$ when $\sigma_{x,l} = 0$. In the rest of the paper, we consider the more interesting situations in which $\sigma_{x,l}, \sigma_{x,f} \neq 0$.

Equation 1.2 implies that the average value of x per leader is higher (lower) than the average value of x per person when the covariance of x and the number of followers is positive (negative). This occurs because $\mu_{x/l}$ is a weighted average of the values of x across individuals in which each individual's weight is equal to the individual's number of followers. When $\sigma_{x,f} > 0$, individuals who have higher values of x also have more followers and thus greater weights, which translate into a higher value of $\mu_{x/l}$. Similarly, equation 1.3 implies that the average value of x per follower is higher (lower) than the average value of x per person when the covariance of x and the number of leaders is positive (negative).

1.3 Implications for seeding

The friendship paradox implies that the following method should be used to select a seeding sample in a symmetric network. First, randomly select an edge in the associated undirected graph. Second, randomly choose one of the vertices associated with the edge for inclusion in the seeding sample. Kramer, Cutler, and Radcliffe (2016) and Kumar et al. (2018) noted that this method yields the average number of friends per seed associated with the friendship paradox.

The generalization of the friendship paradox further implies the following method should be used to select a seeding sample in an asymmetric network. First, randomly select an arc $(i, j) \in A$ in the associated directed graph, G . Second, include node j in the seeding sample when seeding with leaders and node i in the seeding sample when seeding with followers. For undirected seeding, first modify the graph G by replacing directed arcs with undirected edges and removing any duplicate edges between pairs of vertices. Second, randomly select an edge in the modified graph. Third, randomly choose either vertex associated with the edge and include it in the seeding sample. We next obtain expressions for the average number of followers obtained per seed using each of these methods.

Seeding with leaders and followers

Let $x = f$ denote the number of followers. Then, $\mu_x = \mu_f$ is the average number of followers per person, $\sigma_{x,f} = \sigma_f^2$ is the variance in the number of followers, $\sigma_{x,l} = \sigma_{f,l}$ is the covariance of the number of followers and leaders, and $\mu_{x/l} = \mu_{f/l}$ ($\mu_{x/f} = \mu_{f/f}$) is the average number of followers per leader (followers per follower).

Thus, equations 1.2 and 1.3 give

$$\mu_{f/l} = \mu_f + \frac{\sigma_f^2}{\mu_f} \quad (1.5)$$

and

$$\mu_{f/f} = \mu_f + \frac{\sigma_{f,l}}{\mu_l}. \quad (1.6)$$

Equation 1.5 implies that the average number of followers per leader is always greater than the average number of followers per person because $\mu_f, \sigma_f^2 > 0$.

Equation 1.6 implies that the average number of followers per follower is greater than (less than) the average number of followers per person when $\sigma_{f,l}$ is positive (negative). Directed trees are examples of networks for which $\mu_{f/f} < \mu_f$.²

Undirected seeding

For undirected seeding, we construct an undirected graph $H(N, E)$ that has the same set of vertices as the directed graph $G(N, A)$. The difference between H and G is that each directed arc $(i, j) \in A$ is replaced with an undirected edge $\{i, j\} \in E$ and any duplicate edges between pairs of vertices are removed. Formally, $\{i, j\} \in E$ if and only if $(i, j) \in A$ and/or $(j, i) \in A$. We say that two individuals (vertices) are friends when they are connected by an edge in H .

We use Eom and Jo's (2014) result for the undirected graph H to obtain an

²Consider a directed tree with $n \geq 2$ nodes. Suppose the arcs are directed away from the root. Let $t \geq 1$ denote the number of leaders of the root. Both the root and its followers have no followers of followers. All other nodes have one follower and one follower of a follower. Thus, the total number of followers of followers is $(n - t - 1)$ and the average number of followers per follower is $\mu_{f/f} = (n - t - 1)/(n - 1)$. The average number of followers per person is $\mu_f = (n - 1)/n$ because each of the $n - 1$ nodes, excluding the root, has one follower. Thus, $\mu_{f/f} < \mu_f$.

expression for the average number of followers from undirected seeding. The key step is using the variable $x = f$ in graph H , in which f is the number of followers associated with a node in G . Then, the average number of followers per friend, $\mu_{f/fr}$, is equal to the average number of followers per seed obtained using undirected seeding. Substituting $x = f$ in equation 1.1 gives

$$\mu_{f/fr} = \mu_f + \frac{\sigma_{f,fr}}{\mu_{fr}} \quad (1.7)$$

where $\mu_x = \mu_f$ is the mean number of followers per individual in G , μ_{fr} is the average number of friends per individual in H , and $\sigma_{f,fr}$ is the covariance of the number of followers in G and the number of friends in H . The sign of the covariance $\sigma_{f,fr}$ determines whether $\mu_{f/fr}$ is greater or less than μ_f . For instance, the average number of followers per friend is smaller than the average number of followers per person in a “star” network, which is a one-level directed tree in which a single person is a common leader of $n - 1 \geq 2$ followers.³



Figure 1.1: Example of a directed graph G (left) and the associated undirected graph H (right)

Figure 1.1 illustrates the relation between a directed graph G and the associated undirected graph H . Graph G shows that (in reality) information flows only from E to F and not from F to E . Graph H , which is used for undirected seeding, ignores this distinction. Table 1.1 shows that G has 9 leaders, 9 followers,

³Consider a star network with $n \geq 3$ nodes and $n - 1$ arcs. There is one “central” node, say s , and an arc from s to each of the other $n - 1$ nodes. The average number of followers per person is $\mu_f = (n - 1)/n$. Undirected seeding assumes that the central node has $n - 1$ friends. Thus, the average number of followers per friend is $\mu_{f/fr} = (n - 1)/(2(n - 1)) = 1/2$. Thus, $\mu_{f/fr} < \mu_f$.

Table 1.1: Average number of followers per seed obtained by seeding with leaders, seeding with followers, and undirected seeding for the relations shown in Figure 1.1

Person	No. of leaders in G	No. of followers of leaders in G	No. of followers in G	No. of followers of followers in G	No. of friends in H	No. of followers of friends in H
A	4	5	4	5	4	5
B	1	4	1	4	1	4
C	1	4	1	4	1	4
D	1	4	1	4	1	4
E	1	4	2	4	2	4
F	1	2	0	0	1	2
Total	9	23	9	21	10	23
Average	$\mu_l = 9/6$	$\mu_{f/l} = 23/9$	$\mu_f = 9/6$	$\mu_{f/f} = 21/9$	$\mu_{fr} = 10/6$	$\mu_{f/fr} = 23/10$

23 followers of leaders, and 21 followers of followers; H has 10 friends (because F , who is only a follower in G , is considered to be E 's friend in H) and 23 followers of friends. As a result, $\mu_{f/fr} = 23/10 < \mu_{f/f} = 21/9 < \mu_{f/l} = 23/9$.

To better understand this result, consider seeding with leaders. Since F has no followers, it cannot be selected as a leader but F can be selected for undirected seeding (with probability $1/10$). Then, since F has no followers and the probability of selecting each other node decreases, the average number of followers per friend is less than the average number of followers per leader. Now consider seeding with followers. Observe that E is a friend of F but not a follower. Thus, undirected seeding increases the probability of seeding E and decreases the probability of seeding the other nodes. E has two followers so undirected seeding yields two more followers than are obtained by seeding with followers. This increase in the number of followers of friends is not large enough to counterbalance the increase in the number of friends relative to the number of followers. As a result, the average number of followers per friend is smaller than the average number of followers per follower. We conclude that, in this example, ignoring the asymmetry in the relation between E and F reduces the average number of followers per seed.

Next, we consider the general conditions under which each of the three

seeding methods obtains the greatest average number of followers per seed. Random seeding is not considered because, as noted, it is always dominated by seeding with leaders. We emphasize that undirected seeding is different from randomly selecting an arc from the directed graph and then randomly selecting either the leader or the follower for seeding. In that case, the average number of followers per seed would always lie between $\mu_{f/f}$ and $\mu_{f/l}$.

Comparison of seeding methods

Comparing equations 1.5, 1.6 and 1.7 gives the conditions under which each of seeding method obtains the greatest average number of followers per seed. Let

$$k = \max \left\{ \sigma_f^2, \sigma_{f,l}, \frac{\mu_f}{\mu_{fr}} \sigma_{f,fr} \right\}.$$

Then, the maximum number of followers per seed is obtained by seeding with leaders when $k = \sigma_f^2$, seeding with followers when $k = \sigma_{f,l}$, and undirected seeding when $k = \frac{\mu_f}{\mu_{fr}} \sigma_{f,fr}$. Observe that $\frac{\mu_f}{\mu_{fr}} \sigma_{f,fr} = \sigma_{f,z}$ where $z = \frac{\mu_f}{\mu_{fr}} fr$ is the variable obtained by rescaling the number of friends in H by μ_f/μ_{fr} . Since the total number of followers cannot exceed the total number of friends, z is a contraction of the variable fr that depends on the number of additional ties in H compared to G . Thus, the method that has then greatest covariation with the number of followers obtains the greatest average number of followers per seed.

There are six possible orderings of the average number of followers per seed obtained by the three seeding methods. Each is feasible and is illustrated in Figure 1.2. The associated calculations are provided in the Appendix except for the network in Figure 1.2(a), which is identical to the example described in Figure 1.1 and Table 1.1.

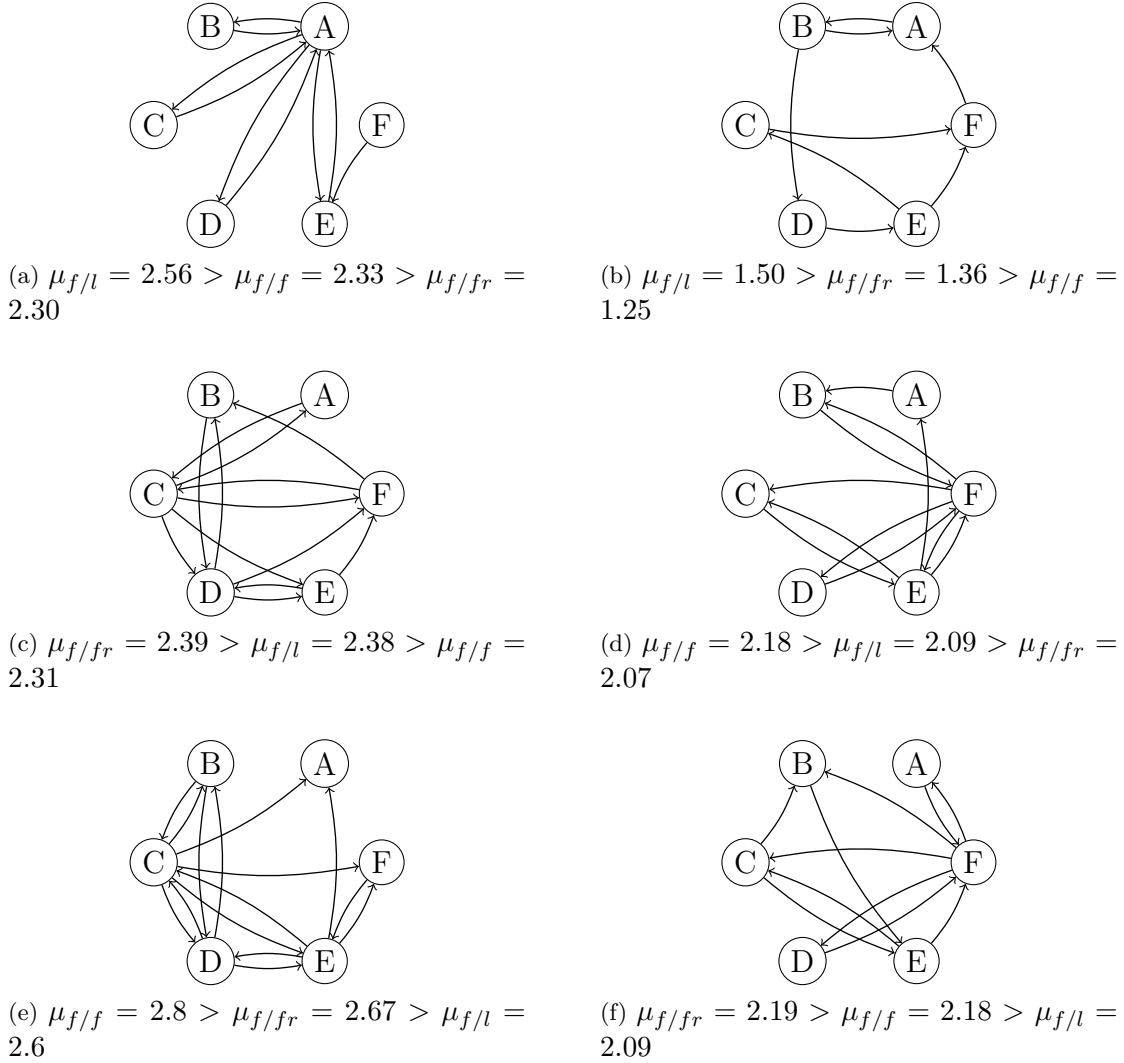


Figure 1.2: Directed graphs illustrating the six possible orderings of the average number of followers per seed

Other measures

The preceding analysis assumes that firms prefer to seed information with individuals who have a large number of followers. We can use Theorem 1 and equation 1.1 to evaluate the seeding methods using other relevant measures. For example, a firm could use a different centrality measure such as eigenvalue centrality, behavioral measures such as the number of messages (tweets and/or retweets) sent by individuals to their followers, or variables such as individual income.

To illustrate, consider a firm that wants to seed individuals who send relatively more messages than others and possesses information on the number of messages, x , that a person in a directed network, G , sends to followers and can compute the values of the covariances $\sigma_{x,f}$, $\sigma_{x,l}$, and $\sigma_{x,z}$ (alternatively, the covariances alone could be provided by the social network). Then, the method with the greatest associated covariance will obtain the greatest average number of messages per seed. Since each of these covariances can be negative, any or all of them can obtain a smaller number of messages per seed than seeding randomly chosen individuals in the network (in contrast, as previously noted, seeding with leaders always obtains more followers per seed than random seeding). Figure 1.3 shows the relation between the average number of followers and the average number of tweets per day sent by Twitter users. The data for the plot were obtained from Sysomos 2009, and the plot suggests a positive covariance between the two variables.⁴ Therefore, seeding with leaders should result not only in a higher number of followers per seed but also in a greater number of messages per seed.

This relation was found empirically by Hodas et al. (2013), who called it the activity paradox. Other measures relevant to seeding should be interesting to pursue in future research. An example is the extent to which information propagates in a social network. Stephen et al. (2017) found that content posted by relatively active users propagates more widely than content posted by lower-activity users. When people who have a relatively large number of followers are also relatively more active, seeding with individuals who have more followers could facilitate greater propagation of information in a network.

⁴The pattern shown in Figure 1.3 only suggests (does not necessarily imply) a positive covariance between the variables because we do not know the number of observations over which the average number of tweets per day was calculated. We analyze Twitter data in section 4 but, unfortunately, do not have data on the number of individual tweets, which is needed to calculate the covariance.

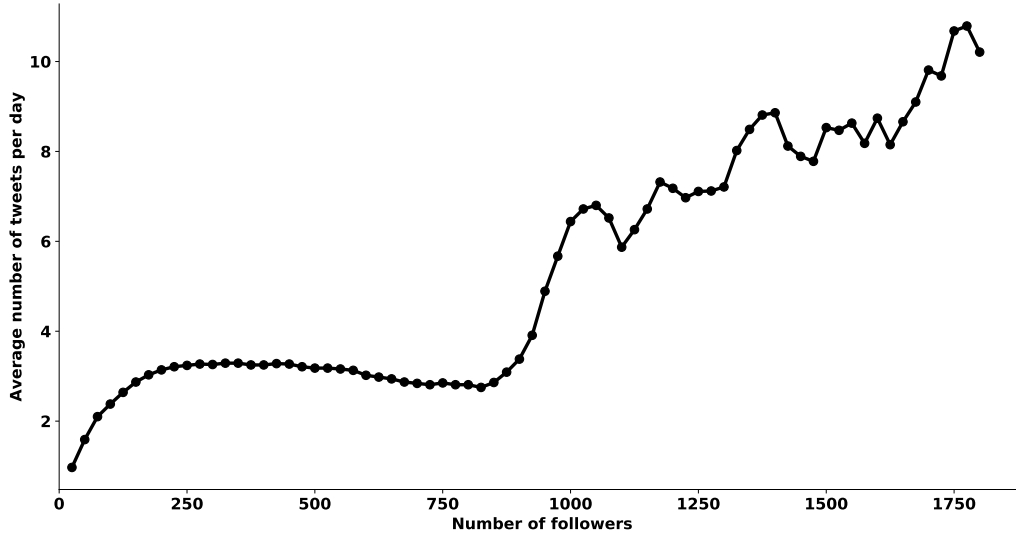


Figure 1.3: Relation between number of followers and average number of tweets per day

1.4 Seeding on Twitter

As noted, Twitter is an asymmetric social network. Firms can seed information in the network by recruiting users offline and/or sending them “promoted tweets” that are sold by Twitter to interested firms as an advertisement channel. Those users can then share the information with their Twitter followers. Karp (2016) reported that nearly 40% of users said they had made purchases as a direct result of tweets from influencers. Gong et al. (2017) found that a media company that tweets about a show on Sina Weibo (“Chinese Twitter”), increases its viewing by 77%; and that viewing increases an additional 33% when an influential person retweets the message. Hodas et al. (2013) analyzed the Twitter data of 5.8 million users and found that the average number of followers per person was smaller than the average number of followers per leader and followers per follower. We extend their analysis to examine the effects of seeding Twitter users.

We obtained data from *networkrepository.com* on 531,000,244 leader-follower

relations among 28,504,110 Twitter users (Rossi and Ahmed, 2015).⁵ The largest network component described by the data accounts for 99.76% of users and 99.88% of leader-follower relations in the sample. We selected a subset of these users by mimicking a screening procedure that Twitter (or another firm) could use before seeding information in the network. First, we removed users who had no leaders because many were likely inactive. This eliminated 85% of the users. Then we removed users who had more than 50,000 leaders and/or followers. Many accounts that have a large number of leaders were likely to be bots, and many accounts with a large number of followers were likely to be celebrities who often charge a substantial fee to promote a message or are unwilling to promote a message.⁶ The remaining 3,653,630 relatively “ordinary” users were used as our sample for seeding.

As Bakshy et al. (2011) noted, seeding “ordinary” users can be more cost effective than seeding celebrities⁷ and, as Watts and Dodds (2007) found, large cascades of influence can be driven by a critical mass of more easily influenced people. Similarly, Cha et al. (2010) found that the most connected users are not necessarily the most influential when it comes to engaging audiences in conversations and spreading messages. Table 1.2 compares the average number of followers across seeding methods for the largest network component with the subnetwork formed by our screening process.

Seeding with leaders in the subnetwork obtained the greatest average number of followers per seed ($\mu_{f/l} = 1,631.42$), followed by seeding with followers ($\mu_{f/f} = 1,479.02$), and undirected seeding ($\mu_{f/fr} = 1,362.22$). All three methods

⁵No descriptive information is available for the data.

⁶Jones (2018) reported that celebrities often charge substantial fees and that even “small” celebrities can charge hundreds, sometimes thousands, of dollars for sending a single tweet. Vosoughi et al. (2018) showed that the pattern of diffusion changes substantially when bots are eliminated.

⁷A limitation of this screening method is that it does not take into account the possibility that seeding a person may become more difficult/expensive when s/he gains more followers.

obtained many more followers per seed than random seeding ($\mu_f = 42.37$). Table 1.2 shows that had we used the full dataset, seeding with followers would have obtained the greatest number of followers per seed, followed by seeding with friends and seeding with leaders. Thus, screening had the effect of reducing the differences in the number of followers per seed obtained using the three methods and increasing the average number of followers obtained by random seeding. In practice, the screening criteria used by a firm affect which seeding method is best. If calculating the average number of followers per seed for each method is difficult, seeding with leaders should be used because it at least guarantees that the number of followers per seed will be greater than the number of followers for a randomly selected seeding sample.

Table 1.2: Average number of followers for different seeding methods using the largest network component and the screened subset of users

	Largest Component	Subnetwork
$ N $	28,434,625	3,653,630
$ A $	530,957,806	154,799,962
$\mu_f = \mu_l$	18.67	42.37
$\mu_{f/l}$	1,252.22	1,631.42
$\mu_{f/f}$	3,932.07	1,479.02
$\mu_{f/fr}$	2,441.41	1,362.22

Figure 1.4 shows the empirical distributions of the total number of followers (F) generated per method for a seeding sample of 1,000 obtained by randomly generating 10,000 samples for each method. The means of the distributions are 42,451.66 for random seeding, 1,595,620.87 for seeding with leaders, 1,449,732.37 for seeding with followers, and 1,338,297.83 for undirected seeding. These values are close to the expected values obtained by multiplying $\mu_f = 42.37$, $\mu_{f/l} = 1,631.42$, $\mu_{f/f} = 1,479.02$ and $\mu_{f/fr} = 1,262.22$ each by 1,000. The estimated standard deviations of the distribution are 8,230.65 for random seeding, 100,039.66 for seeding with leaders, 94,021.65 for seeding with followers and 92,832.144 for

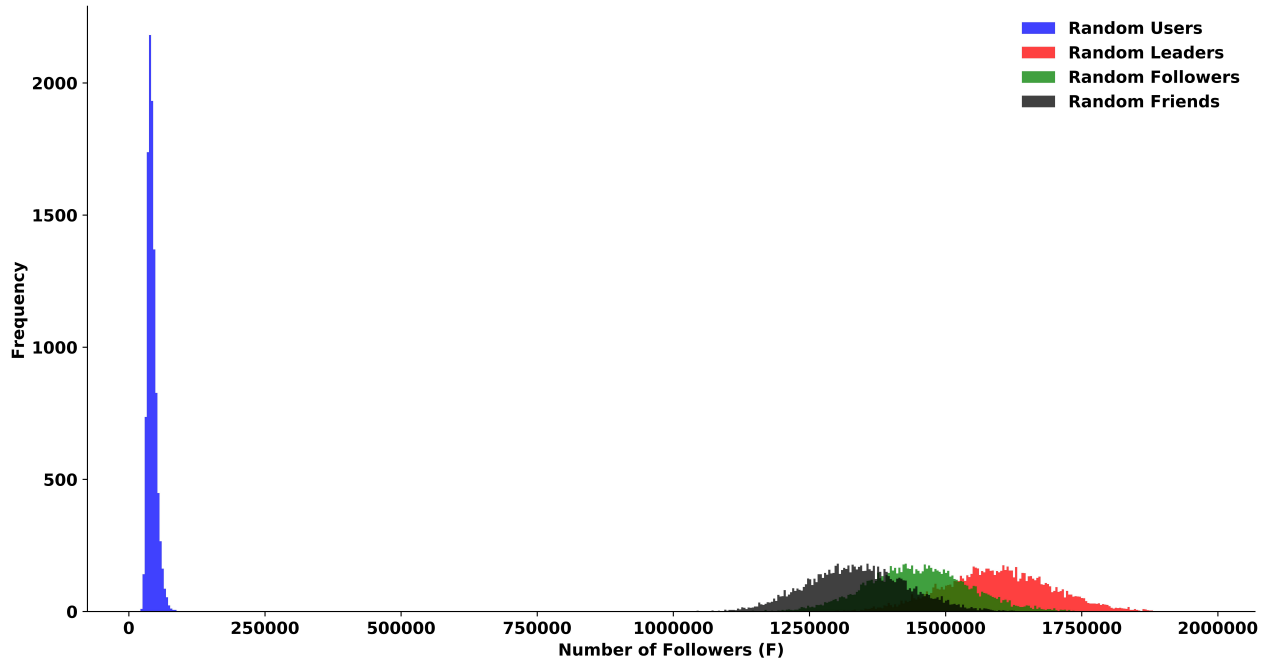


Figure 1.4: Sample distributions of the total number of followers for 1,000 random users, random friends, random leaders, and random followers. Each sample distribution is based on 10,000 random draws.

undirected seeding. Thus, there is little variation in the number of followers from random seeding and approximately equal variation for the other methods.

We use the distributions in Figure 1.4 to estimate the probability that a seeding method will obtain the greatest number of followers in a particular instance. Across 10,000 replications, seeding with leaders obtained the greatest number of followers in 84% of cases, seeding with followers in 14% of cases, and undirected seeding in 2% of cases. Random seeding never obtained the greatest number of followers. Thus, if only a single seeding sample can be obtained, seeding with leaders is recommended. However, if multiple seeding samples can be obtained, the one that generates the greatest number of followers can be selected. For example, using the present analysis, we could obtain 1,970,701 followers by selecting a seeding sample of 1,000 leaders (this corresponds to the rightmost point in the distributions in Figure 1.4).

Reach and duplication

We obtained values of F (total number of followers) by adding the number of followers of each individual in a seeding sample. If two or more seed individuals have the same follower, this method counts some followers more than once. The value of F is thus analogous to the number of gross exposures in advertising and can be decomposed into reach, R , which is the number of unique followers, and duplication, D , which is the average number of individuals in the seeding sample who have the same follower. A larger value of R means that the information will be seeded with a greater number of individuals and a larger value of D means that, on average, a follower can receive information from multiple seed individuals. As the size of a seeding sample increases, so does the probability that two seeds have a common follower. Thus, we expect duplication to increase with the size of the seeding sample.

Figures 1.5 and 1.6 show how the average number of followers and reach and duplication change with the size of the seeding sample. In the figures, the number of individuals in the seeding samples is $s = 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800,$ and $1,000$. Each average is calculated using 10,000 replications.

1. Across seeding samples of different sizes, seeding with leaders obtains a greater average number of followers and greater reach than undirected seeding, seeding with followers, and random seeding. For example, seeding $s = 50$ leaders obtains an average of 81,196.92 followers, which is almost twice the average 42,451.7 followers obtained when seeding 1,000 random users.
2. As Figure 1.5 shows, the number of followers, F , increases approximately linearly with the size of the seeding sample. The rate of growth is fastest for seeding with leaders, followed by seeding with followers and then undirected

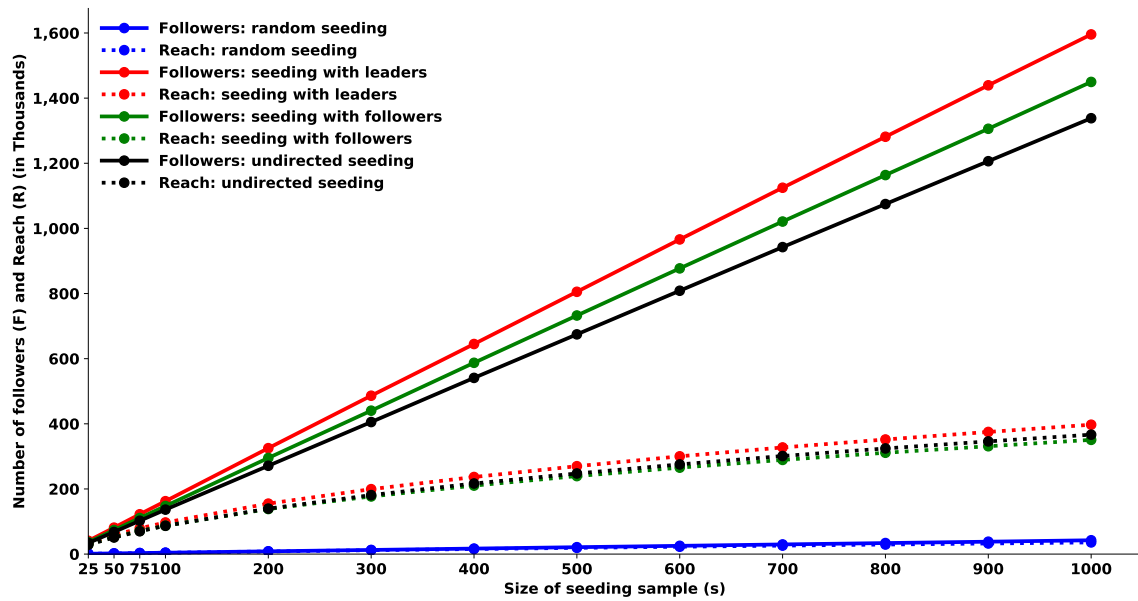


Figure 1.5: Average number of followers (F) and average reach (R) for random users, random followers, and random leaders as a function of the size of the seeding sample (s) across 10,000 runs

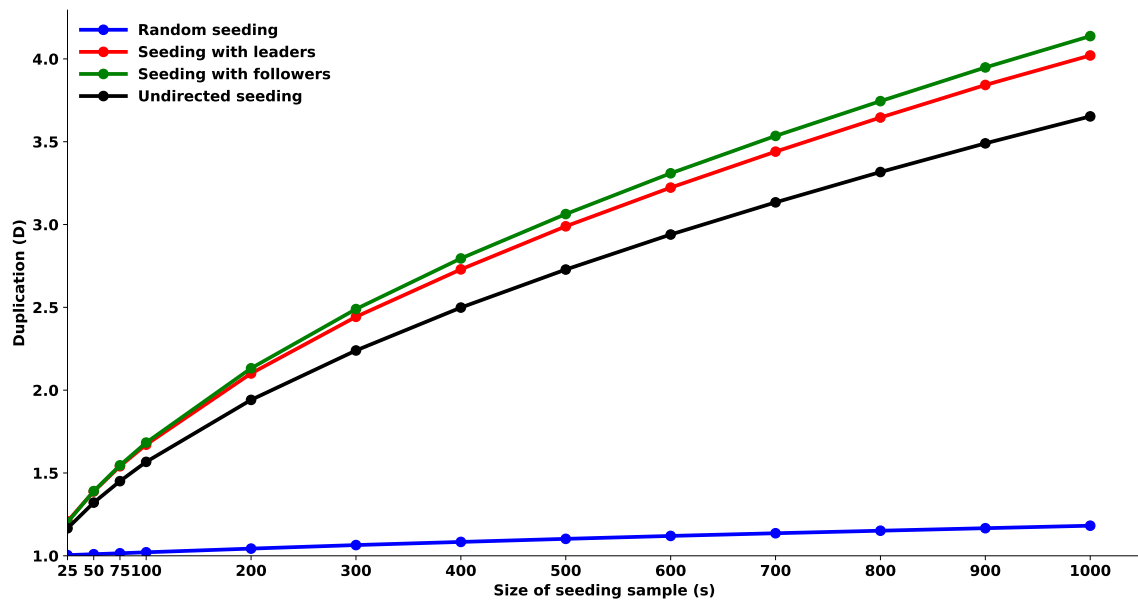


Figure 1.6: Average duplication (D) for randomly selected users, leaders, followers, and friends as a function of the size of the seeding sample (s) across 10,000 runs

seeding. The number of followers grows at a much slower rate for random seeding.

3. Figure 1.5 also shows that the reaches obtained by seeding with leaders, seeding with followers, and undirected seeding grow with the size of the seeding sample at similar rates. The reach for random seeding grows at nearly the same rate as the number of followers; and thus, there is little duplication ($D = 1$) in random seeding.
4. Figure 1.6 shows that seeding with followers leads to greater duplication than seeding with leaders, which obtains higher duplication than undirected seeding. As Watts and Dodds (2007) observed, greater duplication can be useful because receiving a message from multiple people can increase the probability that the receiver will respond to it. Undirected seeding leads to less duplication than seeding with followers, which suggests that it draws a seeding sample from various communities that have few common followers.

1.5 Conclusion

We generalize the friendship paradox to asymmetric relations and examine the implications of the paradox for seeding information in asymmetric social networks. We consider three seeding methods: seeding with leaders, seeding with followers, and undirected seeding, which ignores the distinction between leaders and followers. Only seeding with leaders always obtains a greater number of followers than random seeding. We also characterize the conditions under which each seeding method obtains the greatest number of followers who could potentially receive the seeded information. Our analysis of a large Twitter dataset suggests that all three methods yield substantially more followers than random seeding. Which method obtains the

most followers depends on whether (and how) the seeding sample is restricted using screening criteria. We decompose the total number of followers obtained from the seeding methods to determine their reach (number of unique followers) and duplication (average number of times an individual receives information from more than one seed). The three methods obtained similar reaches, but different levels of duplication. Futures studies could examine other social networks to see if the results are similar. It also could be useful to examine other seeding measures and consider whether the friendship paradox and our proposed generalization to asymmetric relations extend beyond the first level to friends of friends, friends of friends of friends, and so on.

After completion of this thesis, an independent work with a similar contribution was discovered. For reference:

1. Present essay's first online publication (Oct 2018): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248965
2. Independent work's first online publication (May 2019 – to the best of our knowledge): <https://arxiv.org/abs/1905.05286>
3. Independent work's final publication (Feb 2020): <https://www.nature.com/articles/s41467-020-14394-x#MOESM3>

Friendship Paradox Generalizations and Centrality Measures

This essay forms the basis of a paper that is jointly authored with Rajeev Kohli.

2.1 Introduction

Consider an undirected graph in which the vertices represent individuals and the edges represent friendships. Feld (1991) showed that the average degree of the nodes in the graph is no greater than the average degree of the nodes' neighbors. He called this phenomenon the friendship paradox and interpreted it as saying that “your friends have more friends than you” (Feld, 1991). One implication of the paradox suggested and tested by Christakis and Fowler (2010) and Garcia-Herranz et al. (2014) is that a communicable disease such as the flu can appear sooner and spread faster among the friends of a randomly selected group of individuals than among members of the group. Cohen et al. (2003) proposed an immunization strategy for computer networks based on the friendship paradox and Jackson (2019) used the paradox to explain how systematic biases can arise in individual perceptions and affect the formation of social norms. Lerman et al. (2016) related the friendship paradox to a majority illusion in which the preferences of a small group of individuals can be perceived to be the majority view by other people in a social network. Several researchers have examined ways to use the friendship paradox when seeding information in social networks (e.g., Chin et al., 2018; Kumar and Sudhir, 2019; Rubinstein et al., 2015; Singer, 2016).

Let two individuals be each other's k -removed friends when they are connected by a sequence of k intermediate friends. According to the friendship paradox, the expected number of once-removed friends is no smaller than the expected number of zero-removed friends. Kramer et al. (2016) extended this result by showing that the expected number of k -removed friends is no smaller than the expected number of friends when $k \geq 1$ is an odd number.

We consider the relation between the expected numbers of k and $k + 1$

removed friends. First, we show that the expected number of k -removed friends is no greater than the expected number of $k + 1$ removed friends when k is an even number and that no such relation exists when k is an odd number. We then show that, as k increases to infinity, the expected number of k -removed friends converges to the largest eigenvalue of the associated undirected graph regardless of whether k is odd or even.

One implication of these results is that the difference between the expected number of k and $k + 1$ removed friends decreases as k increases. Another is that the friendship paradox is related to beta centrality and eigenvalue centrality. Beta centrality is a weighted sum of an infinite series of the number of $k \geq 0$ removed friends. It converges to eigenvector centrality as the beta parameter approaches the inverse of the limiting value of the expected number of k -removed friends.

We next further generalize these results to directed graphs, which can represent both symmetric and asymmetric relations. An example of an asymmetric relation is an unreciprocated feeling of friendship with another person. A Twitter user who follows another user but is not followed back is in an asymmetric relation. Hierarchical relations, such as ones between parents and their children, are asymmetric in a different way. We may choose to follow or like another but cannot choose to be someone's parent or child. Undirected graphs cannot be used to represent and analyze asymmetric relations; symmetric relations and asymmetric relations can be represented using directed graphs, which ensure that any forward arc between a pair of nodes is accompanied by a backward arc. Therefore, results for asymmetric relations can be used to obtain equivalent results for symmetric relations but not vice versa. Hereafter, we refer to asymmetric relations generically as leader-follower relations and assume that an arc is directed from a follower to a leader. The results of our analysis do not change when the directions of all the arcs

are reversed.

We show that the last person in a randomly selected alternating sequence of $2k + 1$ leaders and followers (followers and leaders) has no fewer followers (leaders) than the last person in a randomly selected alternating sequence of $2k$ followers and leaders (leaders and followers), for all $k \geq 0$. In the simplest special case, which corresponds to $k = 0$, the expected number of followers of a randomly selected leader and the expected number of leaders of a randomly selected follower are no smaller than the expected number of followers and leaders of a randomly selected person. Hodas et al. (2013) identified these relations empirically by analyzing Twitter data. They also found that, on Twitter, the average number of (1) followers of followers and number of (2) leaders of leaders are no smaller than the average number of followers per person (which is always equal to the average number of leaders per person). However, as Hodas et al. (2013) showed, these relations need not always hold.

We show that, as k increases to infinity, the expected number of leaders (followers) of the last person in a randomly selected sequence of $2k$ alternating leaders and followers (followers and leaders) converges to a value proportional to the largest singular value of the associated (asymmetric) adjacency matrix. We also show that this largest singular value is equal to the geometric mean of (1) the expected number of leaders of an alternating sequence of $2k$ leaders and followers and (2) the expected number of followers of an alternating sequence of $2k$ followers and leaders. As the first expected value (number of leaders) increases, the second (number of followers) decreases.

We then introduce beta centrality measures for leaders and followers. These measures become arbitrarily close to the corresponding left and right singular vectors as the beta parameter approaches the inverse of the largest singular value of

the directed graph. Singular vector centralities for asymmetric relations parallel eigenvector centralities for symmetric relations: an individual is an important leader if s/he is followed by important followers and is an important follower if s/he follows important leaders. Bonacich and Lloyd (2015) noted that when an adjacency matrix is a sociometric choice matrix, the singular vectors identify individuals who are popular and individuals who choose popular individuals. Similarly, when a directed graph represents perceived expertise, a singular vector identifies the experts and the other individuals who are adept at identifying experts. These singular vectors are also the limiting values of hub and authority scores obtained by the HITS (hyperlink-induced topic search) algorithm (Kleinberg, 1999). An authority score is a measure of the relevance of a web page to a search query and a hub score is a measure of the extent to which a web page identifies relevant authorities.

The next section generalizes the friendship paradox and describes its relation to beta centrality and eigenvector centrality. The third section further generalizes the results to asymmetric relations and singular vector measures of leader and follower centrality.

2.2 Symmetric Relations

Consider an undirected graph $G(V, E)$ with $n = |V|$ nodes and $m = |E|$ arcs. The vertices of the graph represent individuals and the edges represent friendships between pairs of people. Let A denote the (symmetric) adjacency matrix of the graph with the uv th element $a_{uv} = 1$ when there is an edge between vertices u and v (individuals u and v are friends) and $a_{uv} = 0$ otherwise.

A walk on G is an alternating sequence of vertices and edges of G . For

brevity, we denote a length k walk by a sequence of vertices v_1, \dots, v_{k+1} in which each successive pair of vertices is connected by an edge, $\{v_i, v_{i+1}\} \in E$, for all $i = 1, \dots, k$. The first and last persons in a walk of length k are each other's $k - 1$ removed friends. So, for example, when v_1, v_2 is a walk of length $k = 1$, then v_1 and v_2 are $k - 1 = 0$ removed friends; that is, they are friends. And when v_1, v_2, v_3 is a walk of length $k = 2$, v_1 and v_3 are $k - 1 = 1$ removed friends; that is, they are friends of a common friend. Note that every person is his/her own $2k - 1$ removed friend for all $k \geq 1$. Let w_k denote the number of walks of length $k \geq 0$. By definition, $w_0 = n$: the number of walks of length zero is equal to the number of individuals. Let $e = (1, \dots, 1)'_n$ denote the unit vector with n elements. It is well known that the elements of A^k count the number of walks of length k between pairs of vertices (e.g., Biggs et al., 1993). Equivalently, the uv th element of A^k is equal to the number of different ways in which individuals u and v are each other's $k - 1$ removed friends. The vector $A^k e$ records the number of $k - 1$ removed friends for each individual, and the scalar $w_k = e' A^k e$ counts the total number of $k - 1$ removed friends across all individuals.

Let $d(v)$ denote the degree of vertex $v \in V$. Randomly select a walk of length k in the graph and then select the vertex at either end of the walk with probability of one half. Label the selected vertex X_k . Then, the expected degree of X_k is equal to the expected number of $k - 1$ removed friends of a randomly selected individual. It has the value

$$E[d(X_k)] = \frac{w_{k+1}}{w_k} = \frac{e' A^{k+1} e}{e' A^k e}.$$

Feld's (1991) friendship paradox is equivalent to the relation $E[d(X_0)] \leq E[d(X_1)]$: the expected number of friends is no greater than the expected number of once-removed friends. We generalize this result to show that $E[d(X_k)] \leq E[d(X_{k+1})]$

when k is any non-negative even number $k = 2l$. That is, the expected number of $k - 1$ removed friends is no greater than the expected number of k -removed friends when k is an even number.

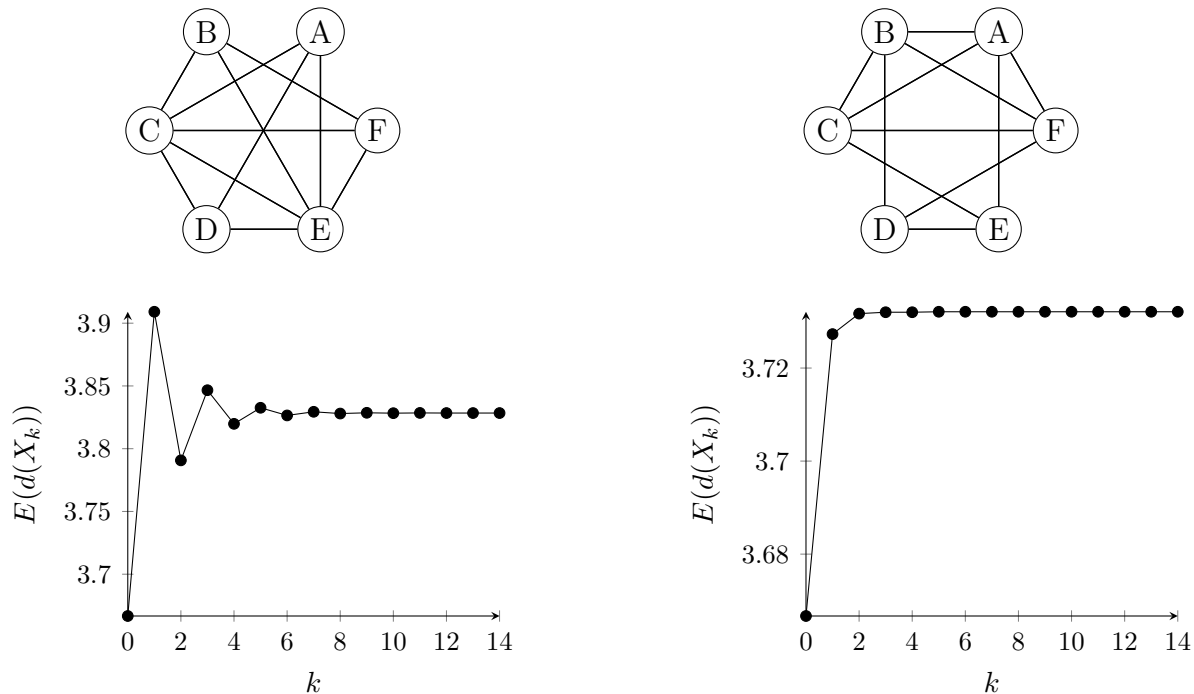


Figure 2.1: Networks for which $E(d(X_1)) > E(d(X_2))$ (left) and $E(d(X_1)) < E(d(X_2))$ (right)

Since A is a symmetric matrix, $A' = A$ and A' denotes the transpose of A . We write $A^{k+1} = A^{2l+1} = A^l A^{l+1}$ and then use the Cauchy-Schwarz inequality to obtain

$$(e' A^{2l+1} e)^2 = (e' A^l A^{l+1} e)^2 \leq (A^l e)' (A^l e) (A^{l+1} e)' (A^{l+1} e) = (e' A^{2l} e) (e' A^{2l+2} e).$$

Dividing both sides of the inequality by $(e' A^{2l} e) (e' A^{2l+1} e)$ yields

$$\frac{e' A^{2l+1} e}{e' A^{2l} e} \leq \frac{e' A^{2l+2} e}{e' A^{2l+1} e}.$$

Since $w_k = e' A^k e$ and $E[d(X_k)] = w_{k+1}/w_k$,

$$E[d(X_k)] \leq E[d(X_{k+1})],$$

which is the desired relation. Thus, the expected number of $k - 1$ removed friends is no greater than the expected number of k -removed friends when $k \geq 2$ is an even number.

There is no similar relation between $E[d(X_k)]$ and $E[d(X_{k+1})]$ when k is an odd number. To see this, consider the example in Figure 2.1. The graph on the left satisfies $E[d(X_1)] > E[d(X_2)]$ and the graph on the right satisfies $E[d(X_1)] < E[d(X_2)]$. In addition, observe that, as the value of k increases, the successive values of $E[d(X_k)]$ oscillate between higher and lower values in the graph on the left and increase monotonically in the graph on the right. Still, in both cases, the difference $E[d(X_k)] - E[d(X_{k-1})]$ decreases toward zero as k increases to infinity. This is because the limiting value of $E[d(X_k)]$ is equal to the largest eigenvalue of A when G is a connected and non-bipartite graph regardless of whether k is odd or even.

Let λ_i denote the i th largest eigenvalue of A and u_i the associated eigenvector. Then,

$$A = \sum_{i=1}^n \lambda_i u_i u_i'.$$

If G is a connected and non-bipartite graph, A is a primitive matrix (König, 1936) and the Perron-Frobenius theorem implies $\lambda_1 > 0$ and $|\lambda_i/\lambda_1| < 1$ for all $i = 2, \dots, n$. Thus, the total number of $k - 1$ removed friends can be expressed as

$$w_k = e' A^k e = \sum_{i=1}^n \lambda_i^k e' u_i u_i' e = \sum_{i=1}^n \lambda_i^k \alpha_i = \lambda_1^k \left(\alpha_1 + \sum_{i=2}^n \left(\frac{\lambda_i}{\lambda_1} \right)^k \alpha_i \right)$$

where $\alpha_i = e'u_i u_i' e = (u_i' e)' (u_i' e) \geq 0$ is a real number.

It follows that

$$\begin{aligned}
\lim_{k \rightarrow \infty} E[d(X_k)] &= \lim_{k \rightarrow \infty} \frac{e' A^k e}{e' A^{k-1} e} \\
&= \lim_{k \rightarrow \infty} \frac{\lambda_1^k \left(\alpha_1 + \sum_{i=2}^N \left(\frac{\lambda_i}{\lambda_1} \right)^k \alpha_i \right)}{\lambda_1^{k-1} \left(\alpha_1 + \sum_{i=2}^N \left(\frac{\lambda_i}{\lambda_1} \right)^{k-1} \alpha_i \right)} \\
&= \lambda_1 \lim_{k \rightarrow \infty} \frac{\alpha_1 + \sum_{i=2}^N \left(\frac{\lambda_i}{\lambda_1} \right)^k \alpha_i}{\alpha_1 + \sum_{i=2}^N \left(\frac{\lambda_i}{\lambda_1} \right)^{k-1} \alpha_i} \\
&= \lambda_1.
\end{aligned}$$

As the value of k increases to infinity, the expected number of $k - 1$ removed friends converges to the largest eigenvalue of A .

We can now relate the friendship paradox to measures of network centrality. Degree centrality is equal to the number of friends. Since $A^k e$ counts each individual's once-removed friends, beta centrality is a weighted sum of the number of $k - 1$ removed friends:

$$c(\beta) = \sum_{k=1}^{\infty} \beta^{k-1} A^k e$$

where β is a positive number. Thus, beta centrality is a weighted sum of the number of k -removed friends across all values of k in which further removed friends are weighted less. In one interpretation, β determines the degree to which status is transmitted from one friend to the next. In another, β is a probability of transmission probability (of a virus or information) from an individual to her/his immediate friend. In the second interpretation, β^{k-1} is the probability of further transmission by a $k - 1$ removed friend and $c(\beta)$ is the expected number of transmissions of a message or virus from different individuals (note that this

interpretation also counts repeated transmissions back to individuals).

It is well known that $c(\beta)$ converges to a finite value only if $\beta < 1/\lambda_1$ and that it approaches eigenvector centrality (the principal eigenvector of A) as β approaches $1/\lambda_1$ from below (Bonacich, 1987, 2007). Equivalently, eigenvector centrality is given by

$$\sum_{k=1}^{\infty} \left(\lim_{l \rightarrow \infty} \left(\frac{e' A^{l-1} e}{e' A^l e} \right) - \epsilon \right)^{k-1} A^k e$$

where $\epsilon > 0$ is an arbitrarily small constant and

$$\lim_{l \rightarrow \infty} \left(\frac{e' A^{l-1} e}{e' A^l e} \right) = \frac{1}{\lambda_1}$$

is the limiting value of the inverse of the expected number of l removed friends.

Eigenvector centrality corresponds to a case of beta centrality in which the degree to which status is transmitted, or the probability of information being transmitted from friend to friend is arbitrarily close to the reciprocal of the limiting value of the expected number of k -removed friends.

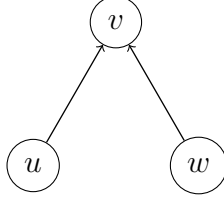
2.3 Asymmetric Relations

We now generalize the preceding results to asymmetric (leader-follower) relations, which are represented by a directed graph $G(V, E)$ with $n = |V|$ nodes and $m = |E|$ arcs. We associate an arc, $(u, v) \in E$, from node u to node v when u is a follower of v (equivalently, v is a leader of u). The leader and follower labels are arbitrary and can be replaced by more suitable ones relevant to the particular context. For example, if the graph shows parent-child relations, an arc can be directed from a parent to a child. If the graph shows Twitter relations, an arc can be directed from

a user to another user s/he follows. And if the graph shows influence relations, an arc can be directed toward a person who influences an individual's decisions.

One difficulty in extending the friendship paradox to asymmetric relations is that, while a person can have leaders and/or followers, leaders do not necessarily have leaders and followers do not necessarily have followers. For example, suppose each of n individuals follows the same person, who follows no one. In that case, there are no leaders of leaders and no followers of followers. However, there are always leaders of followers and followers of leaders. In this example, every follower is the common leader's follower and the common leader is every follower's leader. Consequently, we consider sequences of alternating leaders and followers. Such sequences are key to extending beta centrality and eigenvector centrality to centrality measures for leaders and followers.

Let A denote the adjacency matrix of G with uv th element $a_{uv} = 1$ when there is an arc $(u, v) \in E$ from node u to node v and $a_{uv} = 0$ otherwise. Let $d_{in}(v)$ denote the in-degree (number of followers) and $d_{out}(v)$ the out-degree (number of leaders) of node $v \in V$. The v th element of Ae gives the value of $d_{out}(v)$ and the v th element of $A'e$ gives the value of $d_{in}(v)$. Given an arc $(u, v) \in E$, we define a *forward step* to be a move from node u to node v (from follower to leader) and a *backward step* to be a move from node v to node u (from leader to follower). Define an *alternating walk* of length k to be a sequence of k alternating forward and backward steps (an alternating sequence of leaders and followers). A sequence (v_0, \dots, v_k) of $k + 1$ of successively connected nodes describes an alternating walk of length k when a forward (backward) arc from node v_{j-1} to v_j is followed by a backward (forward) arc from node v_j to v_{j+1} for each $j = 1, \dots, k - 1$. We call this alternating sequence a *forward walk* (*backward walk*) when the first step from v_0 to v_1 is a forward (backward) step. By convention, forward and backward walks of



k	Forward walks	Backward walks
0	$(u), (v), (w)$	$(u), (v), (w)$
1	$(uv), (wv)$	$(vu), (vw)$
2	$(uvu), (uvw), (wvw), (wvu)$	$(vuv), (vuw)$
3	$(uvwv), (uvvw), (wvuw), (wvuuv)$	$(vuvu), (vuvv), (vwvw), (vwvu)$

Figure 2.2: Example illustrating forward and backward walks in a directed graph

length $k = 0$ correspond to the nodes of a directed graph.

Let X_k denote the last node of a randomly selected forward walk of length k and let X'_k denote the last node of a randomly selected backward walk of length k . Figure 2.2 illustrates alternating walks on a graph with $n = 3$ nodes and $m = 2$ arcs. The graph shows that both u and w follow v and that v does not follow anyone. There are two forward walks of length $k = 2$ that end in node u : (wvu) and (uvu) . The walk (wvu) begins with a forward step from w to v and ends with a backward step from v to u and describes the sequence of relations “ w follows v and v leads u .” The walk (uvu) begins with a forward step from u to v and ends with a backward step from v to u . It describes the sequence of relations “ u follows v and v leads u .” Similarly, there are two forward walks of length $k = 2$ that end in node w : (uvw) and (wvw) .

Suppose we randomly select a forward walk of length $k = 2$. Then, $X_2 = u$ or $X_2 = w$ is selected with probability one half. These walks extend to four forward walks of length $k = 3$: $(uvwv)$, $(wvuw)$, $(uvvw)$, and $(wvwv)$. For example, $(uvwv)$ describes the sequence of relations “ u follows v , v leads u , and u follows v .” If we

select any one of these forward walks at random, we always end at node $X_3 = v$ with probability one. Similarly, $X'_2 = v$ with probability one because every backward walk of length $k = 2$ ends in node v , and $X'_3 = u$ or $X'_3 = w$, each with probability one half.

Let α_k denote the number of forward walks of length k and β_k the number of backward walks of length k in G . Then,

$$\alpha_k = e'(AA')^{\lfloor k/2 \rfloor} A^{\lceil k/2 - \lfloor k/2 \rfloor \rceil} e \quad \text{and} \quad \beta_k = e'(A'A)^{\lfloor k/2 \rfloor} A^{\lceil k/2 - \lfloor k/2 \rfloor \rceil} e$$

where $\lfloor * \rfloor \cdot$ denotes the floor function and $\lceil * \rceil \cdot$ denotes the ceiling function. When $k = 2l$ and $l \geq 0$, $\lfloor k/2 \rfloor = \lceil k/2 \rceil = l$ and, thus,

$$\alpha_k = \alpha_{2l} = e'(AA')^l e \quad \text{and} \quad \beta_k = \beta_{2l} = e'(A'A)^l e$$

for all $l \geq 0$. When $k = 2l + 1$ and $l \geq 0$, $\lfloor k/2 \rfloor = l$ and $\lceil k/2 \rceil = l + 1$ and, thus,

$$\alpha_k = \alpha_{2l+1} = e'(AA')^l A e \quad \text{and} \quad \beta_k = \beta_{2l+1} = e'(A'A)^l A' e$$

for all $l \geq 0$. Since $\alpha_{2l+1} = e'(AA')^l A e$ is a scalar, we can write it as

$$\alpha_{2l+1} = e'(AA')^l A e = (e'(AA')^l A e)' = e'(A'A)^l A' e = \beta_{2l+1}.$$

Thus, $\alpha_0 = \beta_0$ and $\alpha_k = \beta_k$ when $k \geq 0$ is odd but not necessarily when it is even. (For symmetric relations, $A = A'$ implies that $\alpha_k = \beta_k$ for both even and odd values of k .)

Consider $k = 0$. Since the elements of the vector Ae list the out-degrees (the numbers of leaders) of the nodes, the X_0 th element of Ae gives the out-degree of the

last node X_0 of a randomly chosen forward walk of length $k = 0$. The expected value of the out-degree of X_0 is equal to the average number of forward walks of length 1 and is given by

$$E(d_{out}(X_0)) = \frac{e' Ae}{e'e} = \frac{\alpha_1}{\alpha_0}$$

where $e' Ae$ is the sum of the out-degrees of all the nodes (the number of forward walks of length $k = 1$) and $e'e = n$ is equal to the number of nodes in the graph. Similarly, since the elements of the vector $A'e$ correspond to the in-degrees of the nodes, the expected value of the in-degree of node X'_0 is given by

$$E(d_{in}(X'_0)) = \frac{e' A'e}{e'e} = \frac{\beta_1}{\beta_0}.$$

Since $\alpha_1 = \beta_1$ and $\alpha_0 = \beta_0$, $E(d_{in}(X'_0)) = E(d_{out}(X_0))$; that is, the expected in-degree is equal to the expected out-degree of the nodes in V (the expected number of leaders is equal to the expected number of followers).

Next, consider random alternating walks of length $k = 1$ in G . First, consider a forward walk chosen randomly from the $\alpha_1 = e' Ae$ forward walks of length 1. Let $X_1 = v$ be the walk's terminal node. The number of forward walks of length 1 that end in v is equal to v 's in-degree which is given by $(A'e)_v$, the v th element of the vector $A'e$. Thus, the probability of randomly selecting a forward walk of length 1 that terminates in node v is equal to $(A'e)_v/(e' Ae)$, the fraction of forward walks of length 1 that end in v . The expected value of the in-degree of X_1 is obtained by weighting the in-degrees of each node v by the probability that $X_1 = v$:

$$E(d_{in}(X_1)) = \sum_{v \in V} \frac{(A'e)_v}{e' Ae} (A'e)_v = \frac{1}{e' Ae} \sum_{v \in V} (A'e)_v (A'e)_v = \frac{e' AA'e}{e' Ae} = \frac{\alpha_2}{\alpha_1}$$

where the second-last equality follows from the observation that multiplying the number of steps that lead to node v by the in-degree of node v is equivalent to

counting the number of forward walks of length 2 in which v is the penultimate node. A parallel calculation shows that the expected value of the out-degree of the terminal node X'_1 of a randomly selected backward walk of length 1 is given by

$$E(d_{out}(X'_1)) = \frac{e' A' A e}{e' A' e} = \frac{\beta_2}{\beta_1}.$$

Lemma 1 generalizes this result to random walks of length $k \geq 1$. We omit the proof, which is straightforward and follows the same reasoning.

Lemma 1. (a) *If $k = 2l$ and $l \geq 0$, then*

$$E(d_{out}(X_{2l})) = \frac{\alpha_{2l+1}}{\alpha_{2l}} \quad \text{and} \quad E(d_{in}(X'_{2l})) = \frac{\beta_{2l+1}}{\beta_{2l}}.$$

(b) *If $k = 2l + 1$ and $l \geq 0$, then*

$$E(d_{in}(X_{2l+1})) = \frac{\alpha_{2l+2}}{\alpha_{2l+1}} \quad \text{and} \quad E(d_{out}(X'_{2l+1})) = \frac{\beta_{2l+2}}{\beta_{2l+1}}.$$

Recall that the expected number of $2k - 1$ removed friends in the symmetric case of is no greater than the expected number of $2k$ removed friends, for all $k \geq 1$. Theorem 2 generalizes this result to leader-follower relations. Proofs of this and the next theorem are given in the Appendix.

Theorem 2. *The expected number of leaders of the last person in a random alternating sequence of $2k$ leaders and followers is no greater than the expected number of leaders of the last person in a random alternating sequence of $2k + 1$ followers and leaders:*

$$E(d_{out}(X_{2k})) \leq E(d_{out}(X'_{2k+1})), \text{ for all } k \geq 0.$$

Similarly, the expected number of followers of the last person in a random alternating sequence of $2k$ followers and leaders is no greater than the expected number of followers of the last person in a random alternating sequence of $2k + 1$ leaders and followers:

$$E(d_{in}(X'_{2k})) \leq E(d_{in}(X_{2k+1})), \text{ for all } k \geq 0.$$

We interpret Theorem 2 for $k = 0$ and $k = 1$. When $k = 0$, the expected number of followers of a randomly chosen person is no greater than the expected number of followers of that person's leader, and the expected number of leaders of a randomly chosen person is no greater than the expected number of leaders of the chosen person's followers. When $k = 1$, the expected number of leaders of a randomly selected follower of a leader (which, for example, can only be u or w in Figure 2.2) is no greater than the expected number of leaders of a randomly selected follower of a leader of a follower (which also can only be u or w in Figure 2.2). Similarly, the expected number of followers of a randomly selected leader of a follower is no greater than the expected number of followers of a randomly selected leader of a follower of a leader.

If each person's follower is also his/her leader, then $d_{out}(X_k) = d_{in}(X_k) = d(X_k)$ and both inequalities in Theorem 2 reduce to $E(d(X_{2k})) \leq E(d(X'_{2k+1}))$ for all $k \geq 0$, which is the result obtained in section 2.2 for symmetric relations. Also, since symmetric relations are special cases of asymmetric relations, the example in Figure 2.1 suffices to show that there is no directional relation between the expected number of leaders (followers) of the last person in a randomly chosen alternating sequence of $2k + 1$ followers and leaders (leaders and followers) and the expected number of leaders (followers) of a randomly

chosen alternating sequence of $2k + 2$ leaders and followers (followers and leaders).

However, regardless of whether k is odd or even, the marginal increase in the number of followers (leaders) obtained by choosing a longer sequence of leaders and followers (followers and leaders) decreases toward zero as the value of k increases. This is a consequence of Theorem 3, which shows that $E(d_{out}(X_k))$ and $E(d_{in}(X'_k))$ converge to values that are proportional to the largest singular value of the adjacency matrix A as k increases to infinity .

Theorem 3. *Let X_k denote the terminal node of a forward walk of length k in G and let X'_k denote the terminal node of a backward walk of length k in G where $k \geq 0$. Let σ_1 be the largest singular value of the adjacency matrix of G . If σ_1 is non-degenerate, then:*

$$\lim_{k \rightarrow \infty} E(d_{out}(X_{2k})) = \lim_{k \rightarrow \infty} E(d_{out}(X'_{2k+1})) = c_{out}\sigma_1$$

$$\lim_{k \rightarrow \infty} E(d_{in}(X'_{2k})) = \lim_{k \rightarrow \infty} E(d_{in}(X_{2k+1})) = c_{in}\sigma_1$$

where c_{out} and c_{in} are constants such that $c_{out} = 1/c_{in}$.

The relation $c_{out} = 1/c_{in}$ in Theorem 3 means that there is a reciprocal relation between the limiting values of the expected number of leaders of a sequence of $2k$ leaders and followers and the expected number of followers of a sequence of $2k + 1$ leaders and followers. The two expected values are equal to the largest singular value when $c_{out} = c_{in} = 1$. Otherwise, as one value increases, the other decreases. Theorem 3 thus implies that

$$\lambda_1 = \sigma_1^2 = \lim_{k \rightarrow \infty} E(d_{out}(X_{2k})E(d_{in}(X_{2k+1})))$$

$$= \lim_{k \rightarrow \infty} E(d_{out}(X'_{2k+1})E(d_{in}(X'_{2k}))).$$

That is, the largest singular value, σ_1 , is equal to the geometric mean of the limiting values of the expected number of leaders of an alternating sequence of $2k$ leaders and followers and the expected number of followers of an alternating sequence of $2k$ followers and leaders.

Let U denote the left singular vector and V the right singular vector associated with σ_1 , which is the largest singular value of A that is assumed to be non-degenerate. These vectors are solutions to the equations $AU = \sigma_1 V$ and $A'V = \sigma_1 U$. Since $A'A$ and AA' have eigenvalue decompositions of $A'A = VD'DV'$ and $AA' = UDD'U'$, the columns of V are the eigenvectors of $A'A$, the columns of U are the eigenvectors of AA' , and the largest eigenvalue of A is $\lambda_1 = \sigma_1^2$. Thus, V corresponds to Kleinberg's (1999) authority (leader) scores and U to hub (follower) scores. Hub and authority scores parallel eigenvector centrality for symmetric relations: a node's authority score is proportional to the sum of the hub scores of its followers and the node's hub score is proportional to the sum of the authority scores of its leaders.

Let

$$l(\beta) = \sum_{k=1}^{\infty} \beta^{k-1} (A'A)^k e$$

and

$$f(\beta) = \sum_{k=1}^{\infty} \beta^{k-1} (AA')^k e$$

where $\beta > 0$ is a constant. We call $l(\beta)$ the beta centrality scores for leaders (authorities) and $f(\beta)$ the beta centrality scores for followers (hubs). These weighted sums converge only if $\beta < 1/\lambda_1 = 1/\sigma_1^2$. As β approaches $1/\lambda_1$ from below, $l(\beta)$ approaches the left singular vector and $f(\beta)$ approaches the right singular vector of A .

Each term in the left and right singular vectors has an interpretation in terms of the expected number of individuals in alternating sequences of leaders and followers and followers and leaders. Bonacich and Lloyd (2015) observed that the singular vectors identify individuals who are popular and those who choose popular individuals when A is a sociometric choice matrix, then . Kleinberg (1999) proposed equivalent hub and authority scores in the context of web searches. Fowler and Jeon (2008) used authority scores to identify the most important precedents of U.S. Supreme Court decisions. Lempel and Moran (2001) developed a related method called SALSA that Twitter employs to recommend other accounts a user might like to follow (Goel et al., 2015). Hub and authority (singular value centrality) scores can also be important in the context of diffusion of innovations. Consider a directed graph in which the nodes represent physicians and an arc is directed from node u to node v if physician u has previously learned about new treatments from physician v . In that case, a physician who has a high authority score plays an important role in disseminating information about new treatments. A physician who has high hub score is someone who can tell you which physicians are important for disseminating such information.

2.4 Conclusion

We generalize the friendship paradox and use the results to relate the friendship paradox to beta centrality and eigenvector centrality. We show that the expected number of k -removed friends is no smaller than the expected number of $k - 1$ removed friends when k is an even number and that this relation does not necessarily exist when k is an odd number. As k increases to infinity, the limiting value of the expected number of k -removed friends converges with the largest

eigenvalue of the associated undirected graph. We interpret beta centrality as a weighted sum of an infinite series of the number of $k \geq 0$ removed friends. It approaches eigenvector centrality when the weighting parameter approaches the inverse of the limiting value of the expected number of k -removed friends.

We further generalize these results to asymmetric relations (e.g., between followers and leaders) represented by directed graphs. We show that the last person in a randomly selected alternating sequence of $2k + 1$ leaders and followers (followers and leaders) has no fewer followers (leaders) than the last person in a randomly selected alternating sequence of $2k$ followers and leaders (leaders and followers) for all $k \geq 0$. As k increases to infinity, the expected number of leaders of the last person in a randomly selected sequence of $2k$ alternating leaders and followers converges to a value proportional to the largest singular value of the associated directed graph. Similarly, the expected number of followers of the last person in a randomly selected sequence of $2k$ alternating followers and leaders converges to a (different) value proportional to the largest singular value of the associated directed graph. We thus show that there is a reciprocal relation between the limiting expected values of leaders and followers. We then generalize beta centrality to asymmetric relations and relate the limiting values of follower and leader beta centrality scores to the singular vectors of the associated directed graph.

The Impact of Auction Houses on Art Valuation

This essay forms the basis of a paper that is jointly authored with Rajeev Kohli and Kamel Jedidi.

3.1 Introduction

About ten thousand years ago, human beings began producing art. Since then, what likely started as a simple expression of creativity has flourished into a major industry. Artists were commissioned, artwork was shown and displayed, and art pieces were sold and auctioned. Indeed, over time, wealthy consumers began investing in art, which was seen as an exclusive form of luxury consumption and as a financial investment (Mandel, 2009). This evolution has accelerated significantly in the past ten years with the total value of artwork worldwide growing 62% and the volume sold growing 34% since 2009. In 2019, the total value reached an estimated \$64.1 billion that was evenly divided between the primary market (e.g., galleries and art dealers) and the secondary market (e.g., auction houses) (McAndrew, 2020). However, art valuations are noted to be skewed. In 2019, only 0.8% of auctioned art pieces were valued at more than \$1 million and yet represented about 55% of total monetary transactions. This unequal distribution is exacerbated by a still imperfect understanding of its drivers. Even major auction houses are at loss to explain why some paintings sell for many times more than expected.

Multiple factors contribute to the difficulty in predicting art prices. First, art pieces are unique works and are rarely sold. Mei and Moses (2002) noted that collectors retain paintings for 28 years on average. In addition, summarizing an artwork based on a few predefined attributes is an arduous task and one that still fails to capture the hedonic aspect of such products. Additionally, pieces vary in terms of significance in art history through their impacts on other artists or paintings; which implies that visual features capture only a small part of the variance. Second, art collector behavior is diverse. Some buy pieces valued under \$1,000 and others are willing to pay more than \$10 million. Their motivations also

vary over a range from conspicuous consumption to strict financial investments. Third, the art market is unregulated (Robertson and Chong, 2008; Meistere, 2018) and data about artwork valuations have been historically difficult to obtain. The primary market represented by galleries and art dealers tends to be secretive regarding prices, which are often discounted or negotiated (Sussman, 2018). It generally has been easier to obtain data regarding the secondary market (auction houses) as information on public auctions has been published online for more than 20 years. However, the information is frequently incomplete and non-standardized, and data related to the auction houses' marketing efforts and bidders' identities, which are likely to affect final valuations, are privately held. Information is also rarely available about private sales.

We focus on auction sales and address three shortcomings in the literature on art valuation. First, existing studies have largely ignored how auction houses directly affect art valuations through their actions and marketing efforts. In particular, in addition to regular auctions held throughout the year, top auction houses (e.g., Christie's, Sotheby's, and Phillips) organize day sales (sometimes divided into morning and afternoon sessions) and evening sales. These marquee events, which are typically associated with much stronger marketing efforts and attract different bidders than regular auctions, represent one mechanism by which auction houses influence prices. Thus, we study how bidders at regular, day and evening events appreciate art through hedonic regressions and assess the causal impact of the auction houses' decisions to sell paintings at evening sales instead of day sales on the final prices of paintings using matching approaches.

Another shortcoming in the literature is the reliance on titles to infer a painting's visual content. Instead, our approach employs deep learning techniques to embed each painting in a low-dimensional representation space. Hence, we can

account for the direct impact of visual features.

The third shortcoming is the literature’s focus on determining the extent of artist influence through their presence in art history books rather than extracting such information at the painting level. Using the paintings representations, we identify the most creative and influential paintings in history by constructing a network of paintings based on visual similarity. The network allows us to quantify the degree of novelty and influence of each work and assess its historical impact.

In particular, having addressed those shortcomings, we can summarize our results as follow:

1. Between 1999-2018, hedonic price indices for regular and day auctions remained relatively stable, leading to overall annualized returns of 3.93% and 4.52%, respectively. In comparison, the hedonic price index for evening auctions grew tremendously, aside from a large drop in 2008, with an average annualized return of 14.33%.
2. Art pieces sold in different types of auctions appreciate differently. For instance, paintings sold in evening auctions tend to be valued for their influence on other paintings while the opposite is true for regular auctions. We interpret this result by noticing that influential paintings sold in regular auctions are likely seen as “mainstream”.
3. By choosing to auction an artwork in an evening sale instead of a day sale, auction houses increase the price of a painting by almost \$6 million on average.

Our aim is to assist multiple stakeholders in the art industry. The tool we create is beneficial to artists and to the primary market because it allows gallery owners to detect promising artists. Additionally, it is beneficial to auction houses, art buyers, and investors as it can help them determine how to allocate their

resources. Finally, by describing which paintings are most influential over time the elements that make a painting especially valuable, our model can be useful to art teachers and to improve art education.

We collected data on every fine art auction conducted in New York City from 1999 through 2018, creating a dataset of more than 140,000 auctions selling paintings made by about 19,500 artists across 11 auction houses. This rich dataset contains information on the artists (e.g., country, birth/death), paintings (e.g., medium used), auctions (e.g., realized price), and digital images of the paintings.

3.2 Literature Review

3.2.1 Art as an Investment

We build on a growing stream of research dedicated to the study of art as an investment. Early attempts to design indices of art returns were agnostic to the paintings' attributes and relied solely on their hammer prices (Stein, 1977; Baumol, 1986). More recent research, akin to studies of housing prices, has delved deeper into understanding the drivers of art valuations using repeat-sales regressions (RSRs) and hedonic pricing. RSR methods involve building datasets for every item sold in an auction at least twice and studying the average rate of return on those items (Anderson, 1974; Goetzmann, 1993; Pesando, 1993; Mei and Moses, 2002). This approach is particularly interesting when some attributes of the items are not observed by the researchers (e.g., the historical context of a painting) or are difficult to encode (e.g., visual features).

By analyzing the evolution of a single painting's valuation over time,

researchers automatically control for differences in attributes that otherwise could influence sale prices. However, significant caveats are associated with this method. First, the scope of these studies is limited to estimating a price index that is used to compare returns from art versus more traditional investments (Chanel et al., 1996). Thus, by foregoing some information related to each transaction, this approach does not allow researchers to generate insights regarding the features that drive the hammer prices and cannot predict future prices. Second, Renneboog and Spaenjers (2013) noted that art pieces tend to be traded infrequently, resulting in a small number of observations. In turn, the small datasets can lead to selection biases depending on the selection procedures used (Mei and Moses, 2002) and to unreliable estimators (Wallace and Meese, 1997). These caveats led researchers to analyze art prices using hedonic regressions.

At its core, a hedonic regression formalizes prices as a function of objective characteristics that implicitly affect buyer utility (Rosen, 1974; Nelson, 1978). Theoretically, this approach allows one to use larger datasets that are not limited to repeated sales and to study both the effect of paintings' characteristics and trends in the paintings' prices through time dummies. However, in practice, researchers have not reached a consensus regarding the observable characteristics to use as predictors (Ashenfelter and Graddy, 2003). In addition, they have incorporated only a few aspects related to visual features of paintings and have applied their approaches to highly specific samples. For instance, Chanel et al. (1996) estimated a price index based on 46 artists and about 1,900 paintings, Collins et al. (2009) focused on Symbolist paintings auctioned between 1990 and 2001, and Galbraith and Hodgson (2018) analyzed paintings from 64 Canadian painters. A notable exception is the study by Renneboog and Spaenjers (2013), which examined more than one million oil paintings, watercolors, and drawings sold from 1957 through 2008 in several marketplaces such as New York and London. Aside from the large

sample size, the authors also included information on the artists and extracted the paintings' topics from the titles. In contrast, the dataset we use focuses exclusively on paintings sold by New York auction houses from 1999 through 2018 and includes images of the paintings, allowing us to study the impact of visual features and creativity more thoroughly.

3.2.2 Computer Vision

Methodologically, our work bridges two approaches used in the marketing literature to extract insights from image data. Liu et al. (2019, 2020) applied supervised deep learning models to analyze raw images and extract visual attributes to answer marketing questions. However, that approach presents two major drawbacks. First, by definition, those models require a specific dependent variable. Second, once the model is trained, the visual aspects extracted could be meaningful only in relation to the dependent variable. Notice that it would be possible to adopt a supervised approach in which we predict painting prices based on their images. However, although this method could potentially offer a strong predictive power, it would not allow us to generate interpretable insights from our results. Instead, our context requires us to work with task-invariant visual features. Dew et al. (2019) developed a semi-supervised generative model (multimodal variational autoencoder) to study different aspects of a brand logo. That method allows one to represent a logo in a low-dimensional representation that is not specific to a given task. However, to do so, the authors designed a feature extraction tool to synthesize a logo into a set of variables that they combined with other brand-specific features. Their strategy is not amenable to analyzing the paintings in our dataset since artworks are not designed to represent a brand or a unique message. Instead, we constructed a variational autoencoder (VAE) that allows us to work with raw images while

extracting task-invariant features.

VAEs (Kingma and Welling, 2013; Rezende et al., 2014) are self-supervised generative models that can be seen as a combination of two neural networks. The first network, the encoder, takes the raw pixel-level image as an input and outputs a vector in a lower dimensional representation space. The second network, the decoder, transforms the vector into an image. The two networks are then trained together such that the image provided by the decoder resembles the original image.

3.2.3 Social Network Analysis

Our study is also relevant to the literature on social network analysis. While the marketing literature has mostly focused on networked human consumers, it is also possible to conceptualize networks of entities such as citations and artworks. For example, Fowler and Jeon (2008) created a network of citations in supreme court cases to determine hub and authority scores (Kleinberg, 1999), Saleh et al. (2016) examined similarities between paintings to create a graph that summarized the influences artists had on each other, and Monechi et al. (2017) constructed a co-occurrence network of user-generated tags between albums that evolved over time to study whether new albums create new connections between tags.

The present study adopts a different approach that consists of deriving a network of art pieces based on their visual similarities and their conception dates. Our methodology is closely related to works by Elgammal and Saleh (2015) and Shrivastava et al. (2017) but differs in significant ways. First, rather than constructing a single network of paintings, we study each painting’s historical context among other paintings by the same artist and among paintings created by others. This difference allows us to analyze the impact of a creator’s artistic journey

on art valuations. Second, we do not aim to summarize an art piece’s creativity as a unique centrality variable per painting and differentiate a painting’s novelty and influence through its in-degrees and out-degrees. Furthermore, we exploit two measures derived from the friendship paradox (Feld, 1991) that incorporate information on the novelty and influence of the paintings to which a work is connected.

Specifically, for a focal painting, we measure (1) whether pieces influenced by it tend to be influenced by many other paintings and (2) whether many pieces were influenced by the same paintings that influenced the focal one. Hence, this methodology allows us to attain a more complete understanding of the historical context of each painting.

In the next section, we describe the dataset we compiled and the feature engineering we undertook to extract visual features and understand the historical context of a painting. We then present the results of our hedonic regression analyses and propensity score matching. Finally, we conclude with a summary of our findings.

3.3 The Dataset

The art market tends to be extremely fragmented and is globalized. Some auction houses such as Sotheby’s and Christie’s deal with large portfolios of artworks and customers while smaller auction houses typically focus on domestic markets.

However, in 2019, the three major auction hubs – the United States, the United Kingdom and China – generated about 84% of total global sales value and the United States’ market share, led predominantly by auction houses based in New

York City, was 37% (McAndrew, 2020). In particular, New York City quickly emerged as the leading center of art activities (dealers, fairs, museums, auction houses) and attracted an influential number of fine art sellers and buyers. Hence, to consider a relatively more homogeneous market, we focus our analysis on sales of paintings by every auction house in New York City. We further reduce our dataset to auctions that occurred from 1999 through 2018 because (1) on average, paintings are sold only once every 28 years (Mei and Moses, 2002), which indicates that most painting sales would be unique; and (2) images of paintings sold before 1999 were difficult to obtain.

Our data collection strategy departs from most previous studies that focused on art movements (Collins et al., 2009) or on establishing a list of artists based on art history resources (Renneboog and Spaenjers, 2013). Those approaches can create incomplete datasets since some artists explicitly reject belonging to a specific category and may underrepresent contemporary artists. Instead, we obtained the entire record of auctions conducted in the leading marketplace and employed several key variables.

After removing art pieces that were not sold and entries for which an image was not readily available, we arrived at a final dataset consisting of 106,363 auction results across 11 auction houses (Bloomsbury Auctions, Bonhams, Christie's, Doyle, Gene Shapiro, Heritage Auctions, Keno Auctions, Phillips, Shapiro Auctions, Sotheby's, and Swann) and 16,902 artists. Upon further investigation, we found that three auction houses had a different focus and had sold less than 0.1% of the paintings in the dataset (at most 108). Therefore, we removed those auction results from the analysis, leaving 8 auction houses.

Each auction observation included data for a diverse set of variables at the auction and painting level. Additionally, we performed some feature engineering to

extend the variables using information about the paintings’ visual features and “network” variables that captured each painting’s historical context.

3.3.1 Dependent Variables

Our primary dependent variable of interest corresponds to the logarithm of the sale price (including the buyer’s premium, which is the fee paid by the buyer to the auction house) of each work. To compare those prices over our time period, we adjusted the nominal values for inflation in 2018 U.S. dollars using the Consumer Price Index.

3.3.2 Auction-level Variables

Several hedonic variables can be attributed to characteristics of the auction:

- Auction House: We introduce a dummy variable for each of the eight auction houses.
- Auction Date: Consistent with earlier studies on the topic (e.g., Renneboog and Spaenjers, 2013), we create one dummy variable for the auction year and one for the month of sale to capture trend and seasonality.
- Auction Description: Auctions are usually accompanied by short titles describing a theme. For instance, on February 12, 2002, Doyle hosted an auction titled “Dogs in Art” and one on May 21, 2002 titled “European and American Paintings and Sculptures.” We capitalize on this information using a set of nine dummy variables (European, American, Old Masters, Contemporary, Modern, Post-War, Impressionist, Latin, and Dogs) as proxies of the movements to which the paintings belong.

3.3.3 Painting-level Variables

- Artist Name: Because of the large number of artists in our sample, we created dummy variables only for artists who had sold more than 100 pieces in the sampled auctions and pooled the remaining artists together. This resulted in 193 non pooled artists who accounted for 27.48% of the artworks.
- Painting Size: Having obtained the heights and widths of the paintings, we computed painting areas to summarize their sizes.
- Attribution: Some paintings in the dataset were not officially attributed to an artist and required a specific designation. Thus, we created a dummy variable for each of the following degrees of attribution: After, Attributed to, Circle of, Follower of, Manner of, School of, and Studio of.
- Authenticity: Since many artists add their signatures, initials, and dates (among other inscriptions), we created dummy variables to describe whether the artist signed the pieces and, if so, how.
- Medium: We created dummy variables to describe the type of paint (Oil, Acrylic, Watercolor, Other) and the type of panel (Canvas, Board, Wood, Other) used by the artist.

3.3.4 Visual Features

As previously discussed, in addition to standard covariates, our dataset includes images of the paintings. We use this unstructured data to summarize the works using a small set of variables that encodes key visual features of the paintings. This overall strategy has been used extensively in studies on computer vision in which raw pixel-level data are used as input in a neural network to classify images or

detect objects. In short, neural networks consist of neurons organized in layers that apply a non-linear transformation to the output of neurons in the previous layer. For classification tasks such as the ImageNet Challenge (Russakovsky et al., 2015) the final layer typically outputs the probability that an image belongs to a given category. Those types of supervised models are not easily applicable to our context. An alternative approach would use the output of a layer of a model trained on a given classification task to summarize the images. However, that approach would also fail to extract the most important visual components of the paintings. We cannot use class probabilities (the output of the final layer) since some paintings do not belong to any of the classes (e.g., abstract paintings). Similarly, since this model is initially trained to predict categories, its intermediate layers would contain information related to that objective rather than information on the overall visual features of the paintings. To overcome those shortcomings, we adopt a self-supervised approach that extracts visual features without relying on the dependent variable.

Variational Autoencoders

Variational autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) function at the intersection of graphical models and deep learning and describe tractable frameworks for estimating deep latent-variable models of data. The central idea is to summarize the data into a lower dimensional representation space of latent variables. The autoencoder is composed of two independent, jointly estimated components: an inference model (encoder) that approximates the posterior distribution of the data over latent variables and a generative model (decoder) that reconstructs the observed data given the latent representation.

Formally, for each $i = 1, \dots, N$ paintings, let \mathbf{x}_i represent observation i and

\mathbf{z}_i its latent representation. The generative model aims to learn the joint distribution, which is parameterized using the vector θ , $p_\theta(\mathbf{x}_i, \mathbf{z}_i)$. The joint distribution can be factorized as $p_\theta(\mathbf{x}_i|\mathbf{z}_i)p(\mathbf{z}_i)$ where $p_\theta(\mathbf{x}_i|\mathbf{z}_i)$ corresponds to the distribution of the observed data conditional on the latent variables – that is, the decoder – with parameters estimated from a deep feed-forward neural network and the prior distribution of \mathbf{z}_i . The inference model focuses on the posterior $p_\theta(\mathbf{z}_i|\mathbf{x}_i)$. Since this distribution is usually intractable (Kingma and Welling, 2019), we approximate it through amortized (shared across all observations) variational inference by a distribution $q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ where the elements of ϕ are typically weights and biases from a deep feed-forward neural network, the encoder. Specifically, to constrain the parameters in ϕ to approximate the true distribution, we minimize the Kullback–Leibler divergence (denoted $\mathcal{D}_{KL}(\cdot||\cdot)$) between $q_\phi(\mathbf{z}_i|\mathbf{x}_i)$ and $p_\theta(\mathbf{z}_i|\mathbf{x}_i)$. In turn, as is common for variational bayesian methods, estimating this latent-variable model is equivalent to maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\theta, \phi) = \sum_{i=1}^N \mathbb{E}_{\mathbf{z}_i \sim q_\phi(\mathbf{z}_i|\mathbf{x}_i)} [\log(p_\theta(\mathbf{x}_i|\mathbf{z}_i))] - \mathcal{D}_{KL}(q_\phi(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)),$$

where the first element in the sum corresponds to the reconstruction error and the second element performs a regularization toward the prior distribution. However, this type of objective function cannot be optimized through stochastic gradient descent; it requires use of a reparameterization trick in which the latent variables \mathbf{z}_i are expressed as transformations of independently sampled random variables, allowing us to optimize over the parameters of the transformations.

Implementation

We aim to extract latent variables that accurately summarize the paintings' images in a lower dimensional representation space. The images are represented using the ubiquitous RGB (red, green, blue) color space in which each pixel is represented by a tuple of three integers ranging from 0 to 255 that each correspond to a shade of red, green, and blue, respectively. After standardization, each painting is represented by a three-dimensional tensor of size 299x299x3. The first two dimensions correspond to height and width (the number of pixels in each direction) and the third corresponds to a decimal value between -1 and 1 for each color channels.

In this context, as described by Kingma and Welling (2013), we assume that the prior distribution of the latent variables is a standard multivariate gaussian and that $p_\theta(\mathbf{x}_i|\mathbf{z}_i)$ follows a multivariate normal distribution with mean $DecoderNeuralNet(\mathbf{z}_i)$ and a diagonal covariance $\lambda\mathbb{I}$. Thus, we approximate the true posterior using the distribution $q_\phi(\mathbf{z}|\mathbf{x}_i) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbb{I})$ where $\boldsymbol{\mu}$ and $\log(\boldsymbol{\sigma})$ are outputs of a deep convolutional neural network, $EncoderNeuralNet(x_i)$. When using the reparameterization trick, the latent variables \mathbf{z}_i equal $\boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}$ where \odot represents the element-wise product and $\boldsymbol{\epsilon}$ is sampled from an isotropic gaussian distribution. In summary, after removing constant terms and multiplying by 2λ , we aim to minimize

$$\tilde{\mathcal{L}}(\theta, \phi) = \sum_{i=1}^N \|\mathbf{x}_i - DecoderNeuralNet(\mathbf{z}_i)\|^2 - \lambda \sum_{j=1}^J (1 + \log(\sigma_{ij}) - \mu_{ij}^2 - \sigma_{ij}^2),$$

where J is the dimension of the latent variables. Specifically, we use $J = 100$ to ensure that the representation space is sufficiently large to capture the most important visual features and we fix $\lambda = 0.00005$ to favor the reconstruction term over the regularization.

We build our *EncoderNeuralNet* on top of an Xception model (Chollet, 2017) pretrained on the ImageNet dataset. This network was built on depthwise separable convolutions that allow us to reduce the number of parameters to optimize while achieving a greater accuracy than is possible with models such as InceptionV3 and ResNet-152. As noted by Yosinski et al. (2014), early layers of neural network models trained on images tend to learn similar basic features while later layers extract aspects that are more specific to the task at hand. Thus, by removing the prediction layer and using a pretrained Xception network, we transfer information to our task to achieve better results with fewer optimization steps than would be needed for a random initialization, leading to higher generalization. Note that, to account for the specificity of the paintings, we allow the entire model to retrain. Having inferred the latent variable, we aim to reconstruct the original images.

Intuitively, the *EncoderNeuralNet* reduces our images from tensors of dimension $299 \times 299 \times 3$ to representation spaces of size 100, and our *DecoderNeuralNet* must reconstruct the images from a vector of size 100. Several upsampling methods have been devised to accomplish this task using nearest neighbor and interpolation approaches. However, those methods are pre-defined and do not allow us to improve the fit of the decoder. Therefore, we instead use deconvolutional layers (Zeiler et al., 2010; Dumoulin and Visin, 2016). While convolutional kernels generate smaller outputs by computing linear combinations of sets of pixels, deconvolutional kernels recover the input by summing overlapping convolutional operations on the output. Thus, as with regular convolutional networks, we can reconstruct an image using translation-invariant patterns. Our final architecture is represented in Figure 1 in the Appendix. The number of training iterations (epochs) was determined using a five-fold cross-validation.

In our framework, we use the latent representations in two ways. First, we

perform a principal component analysis to retain 30 components (selected when the eigenvalue is greater than 1, leading to an explained variance of 59.29%) that further summarize information related to the visual features of the paintings. This step allows us to account for potential multicollinearity in our latent representations and to obtain a more scalable model. Note that this additional step still allows us to interpret each component by examining the paintings that have the highest and lowest loadings on each component, though some visual factors may be hard to interpret. For instance, visual factor 7 emphasizes paintings such as Victor Vasarely’s *OND-III* (1968) and Wayne Thiebaud’s *Diagonal Ridge* (1968), which exhibit predominantly blue shades, while *Mojave V* (1970) by Ludwig Sander and *Filtration (Orange)* (1977) by Julian Stanczak have low scores and are predominantly yellow.¹ Table 3.1 presents the paintings that scored highest and lowest on each factor.

Table 3.1: Visual factors and representative paintings

Visual Factor 1	Highest	George Catlin <i>Pam-a-hó, The Swimmer, One of Black Hawk’s Warriors</i>
	Lowest	Neal Tait <i>Hotel</i>
Visual Factor 2	Highest	Hartung Hans <i>P 1960-288</i>
	Lowest	John French Sloan <i>Landscape, Santa Fe</i>
Visual Factor 3	Highest	Moshe Elazar Caster <i>Untitled</i>
	Lowest	Nick Darmstaedter <i>The French Connection II and the Temple of Doom</i>
Visual Factor 4	Highest	Edward Ruscha <i>Howl</i>
	Lowest	Edward Ruscha <i>City with the Jitters</i>
Visual Factor 5	Highest	Raymond De Botton <i>Venetian Glass</i>
	Lowest	Ruth Pastine <i>Legacy</i>
Visual Factor 6	Highest	Henri Le Sidaner <i>Vue De La Terrasse</i>
	Lowest	Élisabeth Sonrel <i>Scenes from Dante Alighieri’s La Vita Nuova</i>
Visual Factor 7	Highest	Victor Vasarely <i>OND-III</i>
	Lowest	Ludwig Sander <i>Mojave V</i>
Visual Factor 8	Highest	David Burliuk <i>Marussia In Springtime</i>
	Lowest	Nicholas Howey <i>East Of End</i>
Visual Factor 9	Highest	Kenneth Noland <i>Twice Told</i>

¹Due to copyrights, we do not include any images of paintings in this paper.

	Lowest	Victor Bauffe <i>Figures In A Field</i>
Visual Factor 10	Highest	Lisa Yuskavage <i>KK (Portrait Of Kathy Kennedy)</i>
	Lowest	Theodore Robinson <i>Field Of Dandelions</i>
Visual Factor 11	Highest	Vincent Gallo <i>Found In Pirandello'S House</i>
	Lowest	Karl Zerbe <i>Still Life In Chelsea</i>
Visual Factor 12	Highest	Mark Flood <i>Enter</i>
	Lowest	Thomas Downing <i>Untitled</i>
Visual Factor 13	Highest	Eric Freeman <i>Untitled (For Bjorn)</i>
	Lowest	Friedrich Voltz <i>Cattle Watering By Stream Under Darkening Skies</i>
Visual Factor 14	Highest	Ganna Kryvolap <i>Donetsk Horizon</i>
	Lowest	William Pierce Stubbs <i>The Willie Reed</i>
Visual Factor 15	Highest	Pleun Piera <i>Resting Traveller Amongst Classical Architectural Ruins</i>
	Lowest	Jeff Elrod <i>Nobody Sees Like Us (Sepia)</i>
Visual Factor 16	Highest	Martin Kippenberger <i>Untitled</i>
	Lowest	Austrian School <i>Liebestod</i>
Visual Factor 17	Highest	After Jacques Charles Oudry <i>Still Life With Sculpture: A Pair of Paintings</i>
	Lowest	Francesco Guardi <i>Port Scenes With Figures; Ruins</i>
Visual Factor 18	Highest	Claude Venard <i>Flowers On A Table Top</i>
	Lowest	Theodoros Stamos <i>The Door II</i>
Visual Factor 19	Highest	Edward Ruscha <i>Remember And Forget (Standing Screen)</i>
	Lowest	Suzanne Eisendieck <i>A La Lisiere Du Bois</i>
Visual Factor 20	Highest	Yayoi Kusama <i>Infinity-Nets By Black</i>
	Lowest	Carl Vilhelm Holsøe <i>Interior, Light Of Spring</i>
Visual Factor 21	Highest	Joseph Stella <i>Pointillist Abstraction (Flowers)</i>
	Lowest	Vasarely Victor <i>Vega-Ball</i>
Visual Factor 22	Highest	Jonas Wood <i>Mini French Open</i>
	Lowest	William Nelson Copley <i>Untitled (Melone, Reganschirn, Fahne)</i>
Visual Factor 23	Highest	Forest Lee Moses <i>Landscape</i>
	Lowest	Henriette Ronner-Knip <i>Playful Pups</i>
Visual Factor 24	Highest	Richard Pousette-Dart <i>Into The Mirror</i>
	Lowest	Peter Halley <i>Age Of Panic</i>
Visual Factor 25	Highest	Motherwell Robert <i>Málaga [Málaga (Spanish Elegy Series)]</i>
	Lowest	Marie Laurencin <i>Vase De Fleurs Jaunes</i>
Visual Factor 26	Highest	Eberhard Havekost <i>Driver 2</i>
	Lowest	John William Godward <i>At The Garden Shrine, Pompeii</i>
Visual Factor 27	Highest	Wojciech Fangor <i>Untitled (#26)</i>
	Lowest	Jiro Yoshihara <i>Untitled</i>
Visual Factor 28	Highest	Kenny Scharf <i>Green Face</i>

	Lowest	Gao Xingjian <i>La Montagne De Reve (The Dream Mountains)</i>
Visual Factor 29	Highest	Aaron Levy <i>Envy</i>
	Lowest	José Agustin Arrieta <i>Naturaleza Muerta Con Papaya Y Limá</i>
Visual Factor 30	Highest	John Mclaughlin #3
	Lowest	Thomas Struth <i>Mailänder Dom (Fassade)</i>

We use the visual factors to control for the aesthetics of the paintings, but our approach can also be used to investigate the relationship between the visual features of paintings and their prices. We then extract “network” variables to measure how creative and influential each painting is.

3.3.5 Network Variables

Having obtained variables that describe the paintings and their auctions, we next characterize the paintings’ historical importance and artistry. Several researchers have studied the relationship between influence and creativity. For instance, Ding et al. (2009) in a study of influential scholars and Zhang et al. (2016) and Chang et al. (2019) in a study of academic papers used citation networks and conceptualized influence as a consequence of creativity. Several other studies in domains such as innovation (Toubia and Netzer, 2017) and text documents (Toubia, 2019) conceptualized creativity as a tension between familiarity and novelty. However, those kinds of approaches do not necessarily apply to art valuation since paintings’ styles can change dramatically over short periods of time (Kim et al., 2014) and the uniqueness of a painting (high creativity and low influence) can be highly valued. Hence, we study creativity and influence separately. Additionally, deviations within an artist’s work may predict his/her future fame (Stamkou et al., 2018). Therefore, we must separately assess each painting’s overall historical

importance among paintings by the same artist and by other artists. Unfortunately, unlike academic articles and court cases that can be assessed using citations, there is no systematic attribution of influence for paintings. To overcome this difficulty, prior studies have focused on building artist networks based on characteristics such as co-exhibition (Fraiberger et al., 2018) that do not permit studying creativity and influence at the painting level. We build on new approaches that construct networks of art pieces, music albums, and movies based on similarity (Elgammal and Saleh, 2015; Monechi et al., 2017; Shrivastava et al., 2017) to derive the relevant variables.

We start by focusing on evaluating a painting’s impact compared to the work of other artists. We construct a weighted directed network of paintings in which the weights correspond to similarity between two paintings and the direction is temporal. Let $\mathcal{P} = \{1, \dots, u, \dots, N\}$ represent our set of paintings, t_u , the year in which painting u was produced,² and a_u , the artist who painted u . Additionally, we consider the function $s(u, v)$, in which $u, v \in \mathcal{P}$, to measure how similar the two paintings, u and v , are to each other. Of the several functions that could be used, we chose the cosine similarity between pairs of latent representations obtained through our VAE. Let p_x^{other} correspond to the value of the x^{th} percentile of all the similarity values between pairs of paintings. We construct a graph, G^{other} , represented by the adjacency matrix $A^{other} = (a_{uv}^{other})_{u,v \in \mathcal{P}}$ $G(\mathcal{P}, \mathcal{A})$ as follows.

$$a_{uv}^{other} = \begin{cases} 1, & \text{if } t_u \leq t_v \text{ and } a_u \neq a_v \text{ and } s(u, v) \geq p_{99}^{other} \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

In other words, we construct a directed arc from painting u to painting v in G^{other} if and only if u was painted before v by another artist and the similarity

²Some creation dates were missing from our dataset. In those cases, we assumed that the artist painted the piece at midlife to impute the missing values.

$s(u, v)$ belongs to the 99th percentile.³ The latter condition requires that the arc is included only when the two paintings are highly similar to each other. The graph G^{other} allows us to conceptualize four measures of centrality that we can interpret in terms of creativity and influence. First, we consider $d_{out}^{other}(u)$, the out-degree of painting $u \in \mathcal{P}$, which is the number of arcs in G^{other} that originate from u . Intuitively, a high out-degree occurs when a painting is highly similar to many paintings subsequently created, meaning that it has had a strong influence. Similarly, we compute the in-degree, $d_{in}^{other}(u)$, of painting u by summing the number of arcs that end in u . A high in-degree occurs when u is highly similar to many older paintings, indicating that the focal painting is less original and/or creative.

To validate our approach, we ideally would rely on a pre-existing network of paintings known to have influenced each other. Since that information is not available, we review our predictions at the artist level instead. To do so, we construct a network of influence among artists using data from *theartstory.org*, a widely read website (more than 50 million viewers) that compiles information related to a select list of artists. In particular, the artists are associated through an “Influences and Connections” tab with up to five other artists who influenced them and up to five artists they influenced. This information allows us to build an artist network from which we compute in-degrees and out-degrees – how many artists influenced them and how many artists they influenced.

Though each artist’s page can refer to no more than ten names, artists can also be referred to in other artists’ pages. Since G^{other} is defined at the painting level, we averaged the in-degrees and out-degrees of the paintings to obtain predicted degrees for each artist. There were 508 artists present in both datasets, and we removed artists who had equal in and out degrees (since G^{other} is unlikely to

³We used the 90th and 95th percentiles and found no significant differences.

predict equal in and out degrees), resulting in 442 artists. We then classified each artist as mostly influenced when the in-degree was larger than the out-degree and as mostly influential otherwise.

We note several limitations of this analysis. First, the network provided by *theartstory.org* might represent visual, ideological, and/or philosophical influences which would make our comparisons difficult. Additionally, in terms of format, an artist may be connected to very few other artists. In those cases, our representations could be incomplete. However, as shown in Table 3.2, the total hit rate for our two networks was 70.14%, higher than the baseline rate of 62.67%. Furthermore, the χ^2 -statistic is statistically significant, meaning that the null hypothesis that the two quantities are independent is rejected and that our overall strategy appears to be valid.

Table 3.2: Cross tabulation between influence predicted by *theartstory.org* and G^{other} . Elements in the brackets correspond, respectively, to expected value and cell χ^2 .

		G^{other}			
		Mostly...	Influenced	Influential	Row Totals
theartstory.org	Influenced		229 [196.16; 5.50]	48 [80.48; 13.34]	277
	Influential		84 [116.84; 9.23]	81 [48.16; 22.40]	165
Column Totals			313	129	442; $\chi^2 = 50.48$ $p < 1e - 4$

In addition to in-degrees and out-degrees, we extract two variables from G^{other} that allow us to further assess the historical importance of each painting. The in-degrees and out-degrees assess whether a painting is influential and original but do not take into account information about the paintings influenced or paintings that influenced it. We are interested in knowing, for example, whether painting u influenced works that were influenced by many different paintings and whether u was influenced by pieces that influenced many other paintings. This recursive

approach builds on the notions of hubs and authorities (Kleinberg, 1999) by accounting for the “quality” of the nodes to which a painting connects.

Let $(\cdot)_u$ correspond to the u^{th} element of a vector and e be a column vector of size N containing 1s everywhere. We define the following two measures:

$$d_{in/out}(u) = \frac{(A^{other} A^{other'} e)_u}{(A^{other} e)_u} \text{ if } (A^{other} e)_u \neq 0, 0 \text{ otherwise,} \quad (3.2)$$

$$d_{out/in}(u) = \frac{(A^{other'} A^{other} e)_u}{(A^{other'} e)_u} \text{ if } (A^{other'} e)_u \neq 0, 0 \text{ otherwise.} \quad (3.3)$$

In equation 3.2, the denominator corresponds to the number of paintings influenced by u and the numerator computes the total number of paintings that influenced paintings influenced by u . Hence, $d_{in/out}(u)$ corresponds to the average number of influencers of paintings that were influenced by u . Similarly, in equation 3.3, the denominator corresponds to the number of paintings that influenced u and the numerator computes the total number of paintings that were influenced by the paintings that influenced u . In other words, $d_{out/in}(u)$ corresponds to the average number of paintings that share the same influencers as u .

Finally, we also create a network of influence, G^{same} , to measure the extent of influence and originality in each artist’s portfolio. This graph is constructed similarly to G^{other} , but the condition $a_u \neq a_v$ is replaced by $a_u = a_v$. We then consider the in and out degrees of each painting u in G^{same} : $d_{in}^{same}(u)$ and $d_{out}^{same}(u)$. Since this graph is extremely sparse, we do not construct recursive variables for G^{same} .

3.4 Model of Art Valuation

Although art auctions are held regularly throughout the year, some events receive particular attention. A few times a year, the most prestigious auction houses (Christie’s, Sotheby’s, Phillips) organize day and evening sales. These marquee events feature selective artworks chosen carefully by auction specialists and attract seasoned collectors (Kuesel, 2019). For evening sales, “champagne is often served, and bidders (and onlookers) typically arrive in formal attire” (Goldstein, 2012). Day sales often occur the day after an evening sale (Goldstein, 2012).

We determine sale type for the paintings in the dataset by leveraging the auction titles. Titles that include the word evening are classified as evening sales, titles that include day, morning, or afternoon are classified as day sales. All other auctions are called regular. As shown in Figure 3.1, we find that prices in each type of auction are significantly different, indicating systematic differences that could be explained by marketing effort, painting quality, or buyer behaviors.

We start by analyzing how art valuation differs across sale type using hedonic regressions and then estimate the causal effect of the decision to sell a painting at an exclusive evening auction using propensity score matching.

3.4.1 Hedonic Regressions

After conducting an 80-20 train/test split of our dataset for each year, we begin our analysis by analyzing regular, day, and evening auctions separately through hedonic regression models (unless specified, the following empirical discussion relies on in-sample data). Specifically, our model regresses the logarithm of the auction price

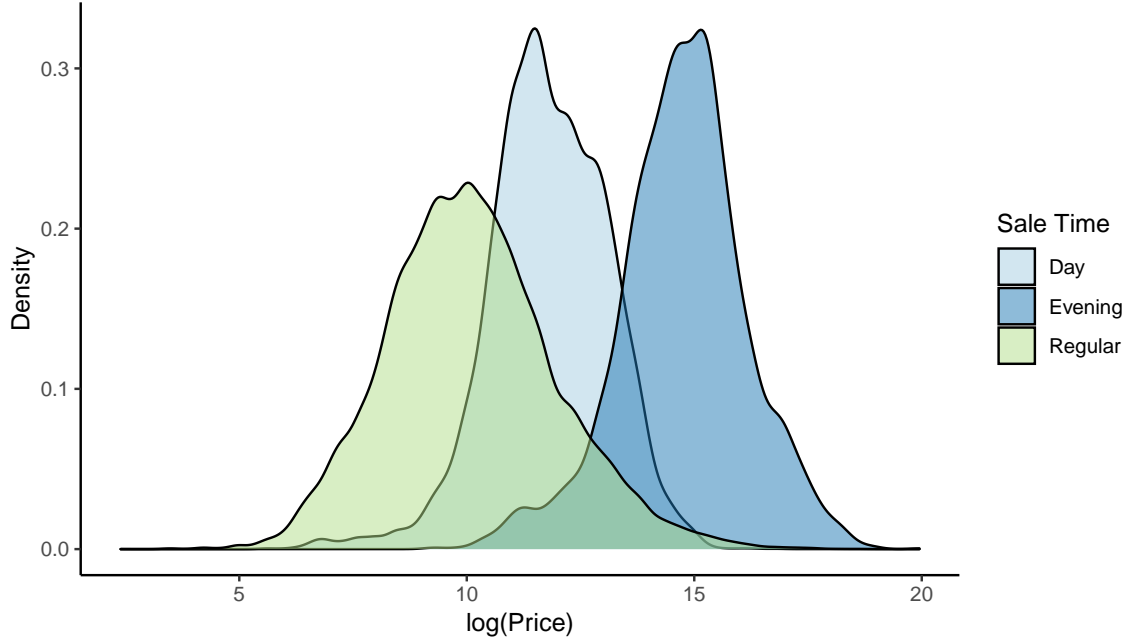


Figure 3.1: Distribution of $\ln(\text{Price})$ for day, evening, and regular auctions

to the predictors as follows:

$$\ln(P_i^{\text{SaleTime}}) = \alpha^{\text{SaleTime}} + \sum_{k=1}^K \beta_k^{\text{SaleTime}} X_{ki}^{\text{SaleTime}} + \epsilon_i^{\text{SaleTime}}$$

where $\text{SaleTime} \in \{\text{Day}, \text{Evening}, \text{Regular}\}$ corresponds to the type of auction, P_i^{SaleTime} is the hammer price (adjusted for inflation) of painting i sold during SaleTime , and X_{ki}^{SaleTime} is the k^{th} characteristic of the painting. For each SaleTime , we estimate (1) a baseline model that contains variables related to the auctions and the paintings, (2) a model that contains the baseline variables and the visual features, and (3) a full model that also includes the network predictors. For ease of readability, we display the results in Tables 6, 7 and 8 in the Appendix. Here we discuss some of the key variables.

Consistent with previous studies of art valuation (e.g. Renneboog and Spaenjers, 2013), we observe that paintings that cannot be fully attributed to an

artist tend to have relatively lower valuations and that use of oil paint tends to have a positive effect on prices. Several of the variables have similar effects on valuation regardless of the type of auction. However, we also observe that the type of auction strongly influences several key components of the paintings' valuations. For instance, while Christie's and Sotheby's are the most important auction houses in our sample both in volume and value, we observe that Christie's under-performs Phillips in regular auctions. In subsequent sections, we focus on returns on art by using the evolution of hedonic price indices and on the interplay between creativity and auction type.

Hedonic Price Indices

In line with previous studies of art valuation, we first evaluate returns on art by establishing hedonic price indices. Our models include time dummy variables, $\beta_t^{SaleTime}$, that correspond to the year $t = 0, \dots, T$ in which a work was sold, which allow us to compute a price index, $H_t^{SaleTime}$, as follows:

$$H_t^{SaleTime} = 100e^{\beta_t^{SaleTime} + \frac{1}{2}(\sigma_t^2 - \sigma_0^2)}$$

where σ_t^2 corresponds to the variance of the residuals of the observations in a specific year t and $\beta_0^{SaleTime} = 0$ (so that our indices are initially standardized at 100) (Triplett, 2004; Silver and Heravi, 2006; Renneboog and Spaenjers, 2013).

Though we have observations beginning in 1999, the first observed day auction is in 2008 and the first observed evening auction is in 2001. These gaps likely occur in part because marquee events do not always follow predictable schedules and plans for such events evolve over time. Christie's and Sotheby's both rescheduled and announced new, initially unplanned, marquee sales in 2020 because of COVID-19.

Furthermore, our data cleaning procedure eliminated some auction results.

Consequently, while each price index is initialized at 100, the first period varies by auction type.

Figure 3.2 represents the evolution of the hedonic price indices over time for the three sale types. We observe a tremendous difference in the price indices. The price indices for both regular and day auctions increased between 2002 and 2007, which was a prolific period for the art industry. The price indices decreased after the financial recession of 2008 and thereafter remained flat. In terms of returns, between 1999 and 2018, annualized returns (the geometric mean of money earned by an investment in that time period) averaged 4.15% for regular auctions and 4.98% for day sales. The price index for evening sales, on the other hand, mostly grew steadily between 2001 and 2018 aside from an expected dip in 2009. After the global recession, the evening auction price index rebounded to pre-crisis levels by about 2015 and the overall annualized return between 1999 and 2018 was 14.76%.

By way of comparison, other studies of art returns that focused on different time periods and did not distinguish auction types found conflicting rates of return. For instance, Mei and Moses (2002) estimated an annualized return rate of 12.81% for the period 1957-1999 using an RSR approach on art auctions. Renneboog and Spaenjers (2013) found significantly lower values of 7.59%, 3.97%, and 5.19% for 1957-1999, 1957-2007, and 1982-2007, respectively using a hedonic pricing approach. Though the source of this discrepancy cannot be causally imputed in our analysis, we have shown that paintings selected for evening auctions have tended to have dramatically stronger rates of return over time. Among the factors that could contribute to this difference are the intrinsic quality of the paintings offered and the types of bidders who tend to participate in evening auctions who could value different aspects of a painting than other bidders. Nonetheless, art investors must

factor in this signal from the auction house before purchasing a painting as an investment.

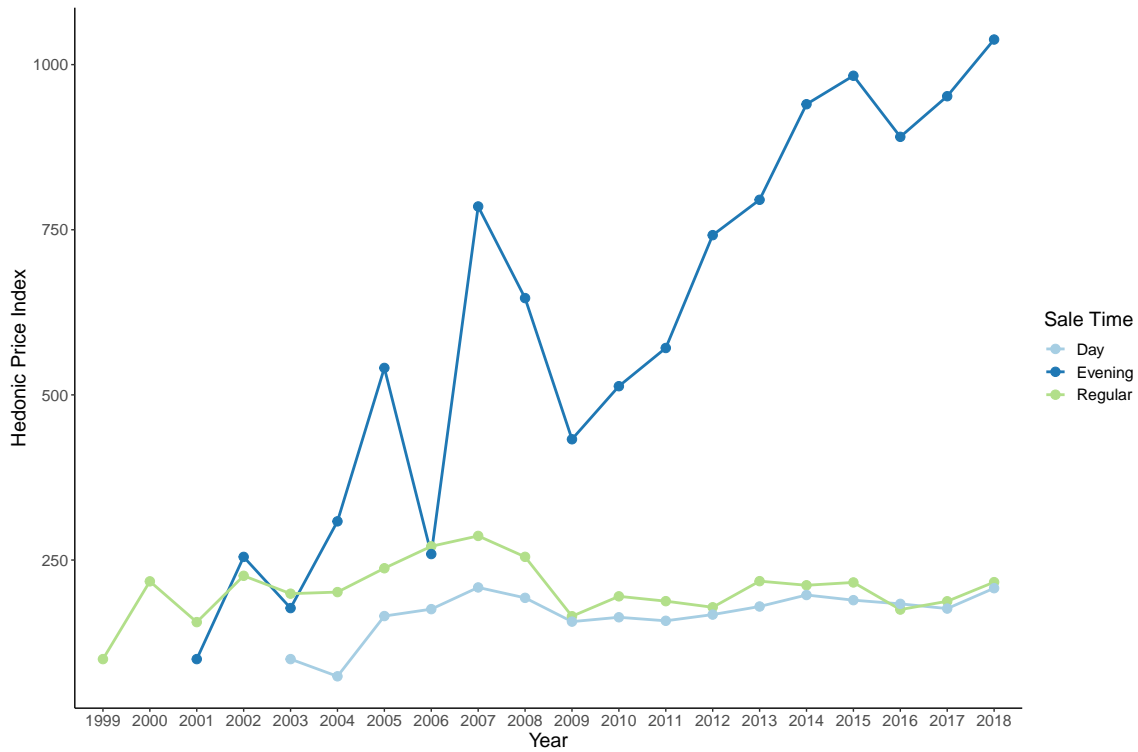


Figure 3.2: Evolution of the hedonic price indices for day, evening, and regular auctions

Creativity

As previously discussed, the bidders who tend to participate in regular, day, and evening auctions are quite different. Therefore, we investigate how paintings' creativity and influence affect their valuations in each type of auction. We start by analyzing any amelioration in the models' performances stemming from including visual features and network metrics in Table 3.3. First, we note that the full model provides a significant improvement in terms of F-statistics for all sale times. In addition, although the improvements in the R^2 and Adjusted R^2 may seem marginal, we observe that the full model also reduces the in-sample and out-sample mean absolute errors (MAEs). In particular, the in-sample MAE reduces from

about \$500 for regular auctions to about \$40,000 in evening auctions. Our full model performs similarly in terms of out-sample MAEs with reductions of about \$700 for regular auctions and more than \$50,000 for evening auctions.

Table 3.3: Model performances for regular, day and evening auctions

	<i>Dependent variable:</i>		
		$\ln(P)$	
	Auctions + Paintings	Baseline + Visual	Full Model
Regular Auctions			
R ²	0.492	0.498	0.499
Adjusted R ²	0.490	0.496	0.497
F-Statistic	264.451*** (df = 260; 70954)	242.521*** (df = 290; 70924)	239.038*** (df = 296; 70918)
In-Sample MAE	184,158.69	183,768.83	183,686.81
Out-Sample MAE	172,654.34	172,010.58	171,947.34
Day Auctions			
R ²	0.461	0.465	0.470
Adjusted R ²	0.451	0.453	0.458
F-Statistic	44.923*** (df = 197; 10347)	39.522*** (df = 227; 10317)	39.188*** (df = 233; 10311)
In-Sample MAE	187,707.32	187,171.54	186,454.67
Out-Sample MAE	167,375.56	167,125.84	166,567.33
Evening Auctions			
R ²	0.472	0.484	0.488
Adjusted R ²	0.448	0.455	0.458
F-Statistic	19.256*** (df = 140; 3011)	16.447*** (df = 170; 2981)	16.098*** (df = 176; 2975)
In-Sample MAE	4,395,924.33	4,364,960.39	4,354,208.72
Out-Sample MAE	5,235,104.11	5,194,797.06	5,181,380.11

Note: MAE corresponds to mean absolute error
Note: *p < 0.1; **p < 0.05; ***p < 0.01

Since the full model represents an improvement in all auction types, we present partial results in Table 3.4 to compare the estimates of the network variables. We start by looking at in-degrees and out-degrees for paintings by the same artist. We observe that the associated coefficients are positive and that

paintings with higher in-degrees and out-degrees receive greater valuations. This indicates that, rather than “seminal work”, paintings that are well established within an artist’s own history tend to generate greater value. This is consistent with previous work that conceptualized creativity as a tension between familiarity and novelty (Toubia, 2019). In addition, in fine arts, artists may make several attempts to generate their own styles early in their careers. Our results indicate that buyers recognize those difficulties and reward paintings that showcase a level of artistic maturity.

In contrast, when considering a painting’s position among pieces by other artists, we observe that d_{in}^{other} has a negative coefficient. That indicates that paintings that are dissimilar from earlier paintings and therefore seen as more creative tend to receive higher valuations. In addition, note the negative coefficient of $d_{out/in}$, the average number of paintings that shared the same influencers as a focal painting, for regular auction. When influencers of a given painting influenced multiple other paintings, valuation of the focal painting diminishes. Hence, to be sold at a higher price, a painting must be dissimilar from other artists’ prior works and be influenced by paintings that had only a minimal impact on other pieces.

However, the effects are not as consistent when considering the influence of a given painting on paintings by others. Relatively influential paintings brought higher prices in evening auctions, but the opposite occurs in regular auctions (the effect is insignificant for day auctions). Additionally, our results show that the coefficient of $d_{in/out}$, the average number of influencers of paintings influenced by a focal painting, is positive for regular auctions and insignificant for evening sales. Hence, the value of a painting sold in a regular auction would increase if paintings it influenced had multiple different influencers. When a focal painting has a low $d_{in/out}$, it indicates that the paintings it influenced were mostly only similar to it.

Thus, we suspect that those paintings were close to a reproduction of our focal painting. Hence, we conjecture that, in lower-tier auctions, paintings with high out-degrees but low $d_{in/out}$ could be seen as “mainstream” work or work that could be easily reproduced and not unique. We leave the study of the tension between ease of reproduction and influence for future research.

Table 3.4: Full model results for regular, day, and evening auctions

	<i>Dependent variable:</i>		
	$\ln(P)$		
	Regular Auctions	Day Auctions	Evening Auctions
Auction variables	[Included]	[Included]	[Included]
Painting variables	[Included]	[Included]	[Included]
Visual factors	[Included]	[Included]	[Included]
d_{out}^{other}	-0.00001*** (0.00001)	0.00002 (0.00002)	0.00001 (0.00003)
d_{in}^{other}	-0.00001** (0.00000)	-0.00004*** (0.00001)	-0.00004*** (0.00002)
d_{out}^{same}	0.008*** (0.003)	0.007* (0.004)	0.003 (0.005)
d_{in}^{same}	0.015*** (0.003)	0.026*** (0.004)	0.011*** (0.004)
$d_{out/in}$	-0.00002** (0.00001)	0.00000 (0.00001)	-0.00001 (0.00003)
$d_{in/out}$	0.00001** (0.00000)	-0.00001* (0.00001)	-0.00002 (0.00002)
Observations	89,014	13,155	3,973
R ²	0.499	0.470	0.488
Adjusted R ²	0.497	0.466	0.458
F-Statistic	239.038*** (df = 296; 70918)	39.188*** (df = 233; 10311)	16.098*** (df = 176; 2975)

Note:

*p < 0.1; **p < 0.05; ***p < 0.01

3.4.2 Propensity Score Matching

The preceding analyses offer interesting insights into the art market at different auction tiers. However, our results are inherently correlational. Notoriously influential paintings have greater odds of being auctioned at Sotheby’s or Christie’s, and several unobservable variables, such as each painting’s bidding process and buyers’ identities, also potentially affect art valuations and could modify our understanding of the market. In particular, we have shown that paintings sold at evening auctions tend to have higher valuations than paintings sold at day and regular auctions and that several differences arise from the type of auction. Due to selection biases, we cannot causally attribute the difference in prices to auction type using hedonic models. Indeed, evening auctions are carefully designed. For instance, art specialists find that some paintings may sell for less than they expect when sold in evening auctions since they would not receive enough attention as other works in the auction (Kuesel, 2019). In addition, auction houses have devised several strategies to convince sellers to retain their services, including third-party guarantees that ensure that at least one bidder will be interested in purchasing the work, leading to “staged” events (Dobrzynski, 2015).

Yet, it is particularly important to estimate the monetary impact of auctioning a painting at an evening sale. In our dataset, that decision is a proxy for marketing efforts implemented by auction houses. Thus, our results allow art specialists to evaluate the performances of auctions and of associated marketing, guide their future decisions, and facilitate the negotiations between sellers and auction houses. Since there is a greater overlap between paintings sold in day and evening auctions (Kuesel, 2019), we compare day and evening sales and thereafter ignore regular auctions. Ideally, to identify a causal impact, we would conduct a

randomized controlled trial in which the paintings that meet the criteria for either an evening or a day auction would be randomly assigned to one of the events. Such an experiment would require randomizing the order in which the paintings are auctioned to mitigate potential spillover effects and overshadowing from some paintings. This approach is obviously unrealistic due to the rarity and importance of these events. Instead, we use propensity score matching (Rosenbaum and Rubin, 1983; Dehejia and Wahba, 1999) to create a composite dataset of paintings that is balanced between the treatment (being auctioned at an evening sale) and control (being auctioned at a day sale) conditions to estimate the treatment effect of evening auctions on prices.

We denote $\tilde{P}_i(W_i)$, the potential price outcome of painting i under treatment W_i which is equal to 1 when i is sold at an evening auction and 0 otherwise. The causal effect of the treatment for painting i is $\tau_i = \tilde{P}_i(1) - \tilde{P}_i(0)$ and, consequently, the average treatment effect is $\tau = \mathbb{E}(\tilde{P}_i(1)) - \mathbb{E}(\tilde{P}_i(0))$. However, that quantity cannot be directly computed since (1) for each unit i , only one of the quantities is observed so the observed price equals $P_i = W_i\tilde{P}_i(1) + (1 - W_i)\tilde{P}_i(0)$; and (2) assignment to an evening auction is not random so the paintings in the control and treatment conditions may belong to different populations.

To circumvent this issue, we use the propensity score that corresponds to the conditional probability of being assigned to the treatment condition given a vector of covariates, $\lambda(x) = P(W = 1|X = x)$, and assume that the treatment assignment is strongly ignorable (Rosenbaum and Rubin, 1983). This assumption entails two conditions. First, we assume *unconfoundedness*, which means that the potential outcomes, $(\tilde{P}_i(1), \tilde{P}_i(0))$, are independent of the treatment assignment W_i given a vector of covariate x . Although this criterion is difficult to satisfy (and test) for paintings, we assume that it is respected by accounting for the predictors we

previously established. Indeed, although every painting is inherently unique, controlling for visual aspects, creativity, and influence in addition to characteristics related to the painting and the auction, we extensively limit the risk of hidden biases. Second, we require *overlap*; for a given x , $0 < \lambda(x) < 1$. This assumption usually can be satisfied by modifying the propensity score model. Rosenbaum and Rubin (1983) showed that, under strong ignorability, the average treatment effect can be estimated by sampling units from the treatment and control groups so that their propensity scores are balanced and measuring the causal effect in the resulting subpopulation.

In practice, our propensity scores are obtained by estimating a logistic model in which the dependent variable is the type of auction and the covariates correspond to all the variables previously discussed in Section 3.3. We use the MatchIt package developed by Ho et al. (2011) for a nearest neighbor algorithm to create a balanced dataset that matches paintings in the treatment group to paintings in the control group. Because of the large number of variables, we report the results of matching in Table 9 in the Appendix.

Our matching procedure created control and treatment groups for which a large majority of the variables were statistically insignificant. However, we note that the conditions were not well balanced at the auction house level and for 11 prominent artists. After performing a t-test, our estimate shows that the predicted average treatment effect of evening auctions is significant and equal to \$5,947,127 ($t = 26.797, df = 3984, p < 2.2e - 16$). Thus, by choosing to sell a painting in an evening auction, auction houses can substantially increase the hammer price of the painting. Note that our causal claim is limited. This effect could be driven by hidden variables we do not observe in our dataset so we cannot identify the underlying mechanism.

We can distinguish three major aspects that could nuance our results. First, although we attempted to control for the quality of a painting by including variables related to its visual aspects, creativity, and influence, characteristics such as texture or techniques used by artists may be relevant. Second, we do not observe auction house marketing efforts, which could be inherently different for day and evening auctions. Finally, evening auctions attract wealthier bidders and seasoned collectors who could either judge the paintings using different reference points (in terms of monetary value) or engage in more aggressive bidding. Thus, we must interpret the causal result as a consequence of systematic differences in the auction houses and bidder behaviors engendered by the decision to sell the art in an evening auction.

3.5 Conclusion and Future Research

Our study focuses on furthering understanding of the art market. To this end, we blend tools from computer vision, social network analysis, and econometrics to account for the impacts of paintings' aesthetics, novelty, and influence on prices of paintings auctioned in New York City from 1999 through 2018. We investigate the impact of a painting's characteristics when auctioned at a regular auction versus at a marquee day or evening auction.

Using hedonic regressions, we find that the evolution of hedonic price indices in each type of auction differ greatly. In particular, during the study period (1999-2018), art sold in evening events achieved an annualized return of 14.33% while art sold in regular and day auctions achieved 3.93% and 4.52%, respectively. We also observe that buyers who attend each type of auction appreciate different elements of paintings. In evening auctions, we find that higher similarity with newer paintings is consistent with the belief that a painting was influential and lead to a

higher valuation; whereas in regular auctions, a higher similarity is correlated with a lower value. We conjecture that in the latter case, high similarity reflects “mainstream” styles of art. We also adopted a causal approach based on propensity score matching and find that the decision to auction a painting during a marquee evening event increases its price by almost \$6 million. To our knowledge, this is the first study to quantify the direct impacts of auction house decisions on prices.

We also acknowledge several limitations of our study. First, although our work provides interesting insights into the importance of creativity in the art world, future work should aim to validate our network measures further, potentially through randomized experiments. Second, our dataset, albeit rich, contains information only about the supply side of auctions and does not address elements regarding the buyers. Hence, our hedonic regression analysis remains correlational, and we cannot fully understand how bidding behavior and art appreciation change across the distribution of prices. Finally, our causal approach remains imperfect, and auction houses may influence prices using a multitude of mechanisms. As additional data becomes available, these limitations can be resolved.

Conclusion

Our first two essays contribute to understanding of social networks by generalizing the friendship paradox to asymmetric relations and to longer relationships. In particular, we show that a two-step approach to random seeding is always beneficial to a company. Furthermore, although we could potentially improve seeding by considering longer relationships, we would witness diminishing returns that might not justify the cost. Finally, we show that the friendship paradox is deeply connected to the notion of centrality in a network.

Our third essay blends tools from deep learning and social network analysis to understand how art is evaluated. In particular, we show that paintings auctioned in evening sales generated overall annualized returns of about 14%, more than three times the returns generated in regular and day auctions. We further show that the decision to sell a painting in a marquee evening event increases the price by an average of about \$6 million. Aside from purely academic interest, understanding what makes art valuable is essential for many stakeholders in the industry. This tool can benefit artists and the primary market by allowing gallery owners to detect promising artists. It also allows auction houses to gain negotiating power by

quantifying the impacts of their marketing efforts on prices. Art buyers can improve their investment decision by adopting a more holistic approach to a painting's valuation that incorporates information about its aesthetics and place in history. Finally, by describing which paintings have been most influential over time and the elements that make paintings valuable, our model can assist art teachers and improve art education.

In conclusion, these three studies marry techniques and approaches from graph theory, social network analysis, deep learning, and statistics. The methods developed in this paper can be further expanded to study other creative productions such as movies and sculptures.

Bibliography

- Anderson, R. C. (1974). Paintings as an investment. *Economic Inquiry*, 12(1):13.
- Ashenfelter, O. and Graddy, K. (2003). Auctions and the price of art. *Journal of Economic Literature*, 41(3):763–787.
- Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone’s an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74.
- Ball, B. and Newman, M. E. (2013). Friendship networks and social status. *Network Science*, 1(1):16–30.
- Baumol, W. J. (1986). Unnatural value: or art investment as floating crap game. *The American Economic Review*, 76(2):10–14.
- Biggs, N., Biggs, N. L., and Norman, B. (1993). *Algebraic graph theory*, volume 67. Cambridge university press.
- Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.
- Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social networks*, 29(4):555–564.
- Bonacich, P. and Lloyd, P. (2015). Eigenvector centrality and structural zeroes and ones: When is a neighbor not a neighbor? *Social Networks*, 43:86–90.
- Carley, K. M. and Krackhardt, D. (1996). Cognitive inconsistencies and non-symmetric friendship. *Social networks*, 18(1):1–27.
- Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P. K., et al. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsn*, 10(10–17):30.
- Chanel, O., Gérard-Varet, L.-A., and Ginsburgh, V. (1996). The relevance of hedonic price indices. *Journal of Cultural Economics*, 20(1):1–24.

- Chang, L. L.-H., Phoa, F. K. H., and Nakano, J. (2019). A new metric for the analysis of the scientific article citation network. *IEEE Access*, 7:132027–132032.
- Chin, A., Eckles, D., and Ugander, J. (2018). Evaluating stochastic seeding strategies in networks. *arXiv preprint arXiv:1809.09561*.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.
- Christakis, N. A. and Fowler, J. H. (2010). Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948.
- Cohen, R., Havlin, S., and Ben-Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical review letters*, 91(24):247901.
- Collins, A., Scorcu, A., and Zanola, R. (2009). Reconsidering hedonic art price indexes. *Economics Letters*, 104(2):57–60.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.
- Dew, R., Ansari, A., and Toubia, O. (2019). Letting logos speak: Leveraging multiview representation learning for data-driven logo design. *Available at SSRN 3406857*.
- Ding, Y., Yan, E., Frazho, A., and Caverlee, J. (2009). Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243.
- Dobrzynski, J. H. (2015). How auction houses orchestrate sales for maximum drama.
- Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.
- Elgammal, A. and Saleh, B. (2015). Quantifying creativity in art networks. *arXiv preprint arXiv:1506.00711*.
- Eom, Y.-H. and Jo, H.-H. (2014). Generalized friendship paradox in complex networks: The case of scientific collaboration. *Scientific reports*, 4(1):1–6.
- Feld, S. L. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.
- Fowler, J. H. and Jeon, S. (2008). The authority of supreme court precedent. *Social networks*, 30(1):16–30.

- Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C., and Barabási, A.-L. (2018). Quantifying reputation and success in art. *Science*, 362(6416):825–829.
- Galbraith, J. W. and Hodgson, D. J. (2018). Econometric fine art valuation by combining hedonic and repeat-sales information. *Econometrics*, 6(3):32.
- Garcia-Herranz, M., Moro, E., Cebrian, M., Christakis, N. A., and Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PloS one*, 9(4):e92413.
- Goel, A., Gupta, P., Sirois, J., Wang, D., Sharma, A., and Gurumurthy, S. (2015). The who-to-follow system at twitter: strategy, algorithms, and revenue impact. *Interfaces*, 45(1):98–107.
- Goetzmann, W. N. (1993). Accounting for taste: Art and the financial markets over three centuries. *The American Economic Review*, 83(5):1370–1376.
- Goldenberg, J., Han, S., Lehmann, D. R., and Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of marketing*, 73(2):1–13.
- Goldstein, A. M. (2012). A beginner’s guide to art auctions.
- Gong, S., Zhang, J., Zhao, P., and Jiang, X. (2017). Tweeting as a marketing tool: A field experiment in the tv industry. *Journal of Marketing Research*, 54(6):833–850.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28.
- Hodas, N. O., Kooti, F., and Lerman, K. (2013). Friendship paradox redux: Your friends are more interesting than you. *arXiv preprint arXiv:1304.3480*.
- Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011). Rejoinder—further reflections on studying social influence in new product diffusion. *Marketing Science*, 30(2):230–232.
- Jackson, M. O. (2019). The friendship paradox and systematic biases in perceptions and social norms. *Journal of Political Economy*, 127(2):777–818.
- Jones, S. (2018). A comparison of the paid ad tweet vs the celebrity tweet.
- Karp, K. (2016). New research: The value of influencers on twitter.
- Kim, D., Son, S.-W., and Jeong, H. (2014). Large-scale quantitative analysis of painting arts. *Scientific reports*, 4:7370.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

- Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.
- König, D. (1936). *theory of finite and infinite graphs: combinatorial topology of complexes*, volume 16. Akademische Verlagsgesellschaft mbh.
- Kramer, J. B., Cutler, J., and Radcliffe, A. (2016). The multistep friendship paradox. *The American Mathematical Monthly*, 123(9):900–908.
- Kuesel, C. (2019). The differences between daytime and evening art sales.
- Kumar, V., Krackhardt, D., and Feld, S. (2018). Network interventions based on iniversity: Leveraging the friendship paradox in unknown network structures. Technical report, Technical report, Working Paper, Yale University.
- Kumar, V. and Sudhir, K. (2019). Can friends seed more buzz and adoption?
- Lempel, R. and Moran, S. (2001). Salsa: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems (TOIS)*, 19(2):131–160.
- Lerman, K., Yan, X., and Wu, X.-Z. (2016). The "majority illusion" in social networks. *PloS one*, 11(2):e0147617.
- Libai, B., Muller, E., and Peres, R. (2013). Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion. *Journal of marketing research*, 50(2):161–176.
- Liu, L., Dzyabura, D., and Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science*, 39(4):669–686.
- Liu, X., Lee, D., and Srinivasan, K. (2019). Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning. *Journal of Marketing Research*, 56(6):918–943.
- Mandel, B. R. (2009). Art as an investment and conspicuous consumption good. *American Economic Review*, 99(4):1653–63.
- McAndrew, C. (2020). The art market 2020. *Art Basel & UBS Report*.
- Mei, J. and Moses, M. (2002). Art as an investment and the underperformance of masterpieces. *American Economic Review*, 92(5):1656–1668.
- Meistere, U. (2018). The art market is increasing in transparency.
- Monechi, B., Gravino, P., Servedio, V. D., Tria, F., and Loreto, V. (2017). Significance and popularity in music production. *Royal Society open science*, 4(7):170433.

- Nelson, J. P. (1978). Residential choice, hedonic prices, and the demand for urban air quality. *Journal of urban Economics*, 5(3):357–369.
- Pesando, J. E. (1993). Art as an investment: The market for modern prints. *The American Economic Review*, pages 1075–1089.
- Renneboog, L. and Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. *Management Science*, 59(1):36–53.
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.
- Robertson, I. A. and Chong, D. (2008). *The art business*. Routledge.
- Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rossi, R. and Ahmed, N. (2015). The network data repository with interactive graph analytics and visualization. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- Rubinstein, A., Seeman, L., and Singer, Y. (2015). Approximability of adaptive seeding under knapsack constraints. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 797–814.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Saleh, B., Abe, K., Arora, R. S., and Elgammal, A. (2016). Toward automated discovery of artistic influence. *Multimedia Tools and Applications*, 75(7):3565–3591.
- Shrivastava, D., CG, S. A., Laha, A., and Sankaranarayanan, K. (2017). A machine learning approach for evaluating creative artifacts. *arXiv preprint arXiv:1707.05499*.
- Silver, M. and Heravi, S. (2006). *Why elementary price index number formulas differ: price dispersion and product heterogeneity*. Number 6-174. International Monetary Fund.
- Singer, Y. (2016). Influence maximization through adaptive seeding. *ACM SIGecom Exchanges*, 15(1):32–59.

- Stankou, E., van Kleef, G. A., and Homan, A. C. (2018). The art of influence: When and why deviant artists gain impact. *Journal of personality and social psychology*, 115(2):276.
- Stein, J. P. (1977). The monetary appreciation of paintings. *Journal of political Economy*, 85(5):1021–1035.
- Stephen, A. T., Dover, Y., Muchnik, L., and Goldenberg, J. (2017). Pump it out! the effect of transmitter activity on content propagation in social media. *The Effect of Transmitter Activity on Content Propagation in Social Media (January 1, 2017)*. Saïd Business School WP, 1.
- Sussman, A. L. (2018). Should you expect a discount when buying a work of art?
- Sysomos (2009). Inside twitter: An in-depth look inside the twitter world.
- Toubia, O. (2019). A poisson factorization topic model for the study of creative documents (and their summaries). *Available at SSRN 3334028*.
- Toubia, O. and Netzer, O. (2017). Idea generation, creativity, and prototypicality. *Marketing Science*, 36(1):1–20.
- Triplett, J. (2004). Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products.
- Tucker, C. (2008). Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12):2024–2038.
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Wallace, N. E. and Meese, R. A. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, 14(1-2):51–73.
- Watts, D. J. and Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458.
- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.
- Zeiler, M. D., Krishnan, D., Taylor, G. W., and Fergus, R. (2010). Deconvolutional networks. In *2010 IEEE Computer Society Conference on computer vision and pattern recognition*, pages 2528–2535. IEEE.
- Zhang, S., Zhao, D., Cheng, R., Cheng, J., and Wang, H. (2016). Finding influential papers in citation networks. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 658–662. IEEE.

Appendix A: Essay 1

A1: Proof of Theorem 1

Recall that each node of the directed graph $G(N, A)$ corresponds to an individual. Let $a_{ij} = 1$ if there is an arc from node i to node j (that is, if individual i follows individual j) and $a_{ij} = 0$ otherwise. Let person i have $L_i \geq 0$ leaders and $F_i \geq 0$ followers. Then

$$L_i = \sum_{j=1}^n a_{ij}$$

and

$$F_j = \sum_{i=1}^n a_{ij}.$$

The total number of leaders and followers is

$$m = \sum_{i=1}^n L_i = \sum_{j=1}^n F_j$$

and the average number of leaders and followers per person is

$$\mu_l = \mu_f = m/n,$$

where $n = |N|$ is the total number of individuals (nodes).

Let x denote a variable of interest and x_i its value for node (individual) $i \in N$. Let

$$\mu_x = \frac{1}{n} \sum_{i=1}^n x_i$$

denote the average value of x per individual. The sum of the values of x across

individual i 's leaders is

$$\sum_{j=1}^n a_{ij}x_j$$

and the sum of the value of x across all the leaders of all the individuals is

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij}x_j = \sum_{j=1}^n F_j x_j.$$

Thus, the average value of x per leader is

$$\mu_{x/l} = \frac{1}{m} \sum_{j=1}^n F_j x_j = \frac{1}{n\mu_f} \sum_{j=1}^n F_j x_j.$$

Let $\sigma_{x,f}$ denote the covariance of x and the number of followers. Then

$$\sigma_{x,f} = \frac{1}{n} \sum_{j=1}^n (F_j - \mu_f)(x_j - \mu_x) = \frac{1}{n} \left(\sum_{j=1}^n F_j x_j - \mu_x \sum_{j=1}^n F_j - \mu_f \sum_{j=1}^n x_j + n\mu_f \mu_x \right).$$

Substituting

$$\sum_{j=1}^n F_j = n\mu_f, \quad \sum_{j=1}^n x_j = n\mu_x$$

on the right-hand side and rearranging terms gives

$$\mu_{x/l} = \frac{1}{n\mu_f} \sum_{j=1}^n F_j x_j = \frac{\sigma_{x,f} + \mu_f \mu_x}{\mu_f} = \mu_x + \frac{\sigma_{x,f}}{\mu_f},$$

which is the desired expression for the average value of x per leader.

Next, we derive the expression for the average value of x per follower. We note that the sum of the values of x across individual j 's followers is

$$\sum_{i=1}^n a_{ij}x_i$$

and the sum of the value of x across all the followers of all the individuals is

$$\sum_{j=1}^n \sum_{i=1}^n a_{ij}x_i = \sum_{i=1}^n L_i x_i.$$

Thus, the average value of x per follower is

$$\mu_{x/f} = \frac{1}{m} \sum_{i=1}^n L_i x_i = \frac{1}{n\mu_l} \sum_{i=1}^n L_i x_i.$$

Let $\sigma_{x,l}$ denote the covariance of x and the number of leaders. Then

$$\sigma_{x,l} = \frac{1}{n} \sum_{i=1}^n (L_i - \mu_l)(x_i - \mu_x) = \frac{1}{n} \left(\sum_{i=1}^n L_i x_i - \mu_x \sum_{i=1}^n L_i - \mu_l \sum_{i=1}^n x_i + n\mu_l \mu_x \right).$$

Substituting

$$\sum_{i=1}^n L_i = n\mu_l, \quad \sum_{i=1}^n x_i = n\mu_x$$

on the right-hand side and rearranging terms gives

$$\mu_{x/f} = \frac{1}{n\mu_l} \sum_{i=1}^n L_i x_i = \frac{\sigma_{x,l} + \mu_l \mu_x}{\mu_l} = \mu_x + \frac{\sigma_{x,l}}{\mu_l},$$

which is the desired expression for the average value of x per follower. \square

A2: Calculations of $\mu_{f/l}$, $\mu_{f/f}$ and $\mu_{f/fr}$ for the graphs in Figure 1.2

Table 1: Panel (b): $\mu_{f/l} > \mu_{f/fr} > \mu_{f/f}$

Person	No. of leaders	No. of followers of leaders	No. of followers	No. of followers of followers	No. of friends	No. of followers of friends
A	1	1	2	3	2	3
B	2	3	1	2	2	3
C	1	2	1	1	2	3
D	1	1	1	1	2	2
E	2	3	1	1	3	4
F	1	2	2	2	3	4
Total	8	12	8	10	14	19
Average	$\mu_l = 8/6$	$\mu_{f/l} = 12/8$	$\mu_f = 8/6$	$\mu_{f/f} = 10/8$	$\mu_{fr} = 14/6$	$\mu_{f/fr} = 19/14$

Table 2: Panel (c): $\mu_{f/fr} > \mu_{f/l} > \mu_{f/f}$

Person	No. of leaders	No. of followers of leaders	No. of followers	No. of followers of followers	No. of friends	No. of followers of friends
A	1	2	1	2	1	2
B	1	3	2	6	2	6
C	4	9	2	4	4	9
D	3	7	3	6	4	9
E	2	6	2	5	3	8
F	2	4	3	7	4	9
Total	13	31	13	30	18	43
Average	$\mu_l = 13/6$	$\mu_{f/l} = 31/13$	$\mu_f = 13/6$	$\mu_{f/f} = 30/13$	$\mu_{fr} = 18/6$	$\mu_{f/fr} = 43/18$

Table 3: Panel (d): $\mu_{f/f} > \mu_{f/l} > \mu_{f/fr}$

Person	No. of leaders	No. of followers of leaders	No. of followers	No. of followers of followers	No. of friends	No. of followers of friends
A	1	2	1	2	2	4
B	1	3	2	4	2	4
C	1	2	2	5	2	5
D	1	3	1	3	1	3
E	3	6	2	5	3	6
F	4	7	3	5	4	7
Total	11	23	11	24	14	29
Average	$\mu_l = 11/6$	$\mu_{f/l} = 23/11$	$\mu_f = 11/6$	$\mu_{f/f} = 24/11$	$\mu_{fr} = 14/6$	$\mu_{f/fr} = 29/14$

Table 4: Panel (e): $\mu_{f/f} > \mu_{f/fr} > \mu_{f/l}$

Person	No. of leaders	No. of followers of leaders	No. of followers	No. of followers of followers	No. of friends	No. of followers of friends
A	0	0	2	6	2	6
B	2	6	2	6	2	6
C	5	12	3	8	5	12
D	3	8	3	8	3	8
E	4	10	3	8	4	10
F	1	3	2	6	2	6
Total	15	39	15	42	18	48
Average	$\mu_l = 15/6$	$\mu_{f/l} = 39/15$	$\mu_f = 15/6$	$\mu_{f/f} = 42/15$	$\mu_{fr} = 18/6$	$\mu_{f/fr} = 48/18$

Table 5: Panel (f): $\mu_{f/fr} > \mu_{f/f} > \mu_{f/l}$

Person	No. of leaders	No. of followers of leaders	No. of followers	No. of followers of followers	No. of friends	No. of followers of friends
A	1	3	1	3	1	3
B	1	2	2	5	3	7
C	2	4	2	5	3	7
D	1	3	1	3	1	3
E	2	5	2	4	3	7
F	4	6	3	4	5	8
Total	11	23	11	24	16	35
Average	$\mu_l = 11/6$	$\mu_{f/l} = 23/11$	$\mu_f = 11/6$	$\mu_{f/f} = 24/11$	$\mu_{fr} = 16/6$	$\mu_{f/fr} = 35/16$

Appendix B: Essay 2

B1: Proof of Theorem 1

Recall that the number of forward walks of length $2k + 1$ is

$$\alpha_{2k+1} = e'(AA')^k Ae = e'(AA')^{\lceil k/2 \rceil} (AA')^{\lfloor k/2 \rfloor} Ae.$$

From the Cauchy-Schwarz inequality,

$$\alpha_{2k+1}^2 = (e'(AA')^k Ae)^2 \leq (e'(AA')^{2\lceil k/2 \rceil} e)(e'A'(AA')^{2\lfloor k/2 \rfloor} Ae),$$

which can be written as

$$(e'(AA')^k Ae)^2 \leq (e'(AA')^{2\lceil k/2 \rceil} e)(e'(A'A)^{2\lfloor k/2 \rfloor + 1} e) \quad (1)$$

because $e'A'(AA')^{2\lfloor k/2 \rfloor} Ae = e'(A'A)^{2\lfloor k/2 \rfloor + 1} e$. Similarly, the number of backward walks of length $2k + 1$ is

$$\beta_{2k+1} = e'(A'A)^k A'e = e'(A'A)^{\lceil k/2 \rceil} (A'A)^{\lfloor k/2 \rfloor} A'e.$$

From the Cauchy-Schwarz inequality,

$$\beta_{2k+1}^2 = (e'(A'A)^k A'e)^2 \leq (e'(A'A)^{2\lceil k/2 \rceil} e)(e'A(A'A)^{2\lfloor k/2 \rfloor} A'e),$$

which can be written as

$$(e'(A'A)^k A'e)^2 \leq (e'(A'A)^{2\lceil k/2 \rceil} e)(e'(AA')^{2\lfloor k/2 \rfloor + 1} e) \quad (2)$$

because $e'A(A'A)^{2\lfloor k/2 \rfloor} A'e = e'(AA')^{2\lfloor k/2 \rfloor + 1} e$. We use equations (1) and (2) to prove the theorem separately for odd and even values of k .

(1) If k is odd, then we can rewrite equations (1) and (2) in the form

$$\frac{e'(AA')^k Ae}{e'(A'A)^{2\lfloor k/2 \rfloor + 1} e} \leq \frac{e'(AA')^{2\lceil k/2 \rceil} e}{e'(AA')^k Ae}, \quad (1a)$$

$$\frac{e'(A'A)^k A'e}{e'(AA')^{2\lfloor k/2 \rfloor + 1} e} \leq \frac{e'(A'A)^{2\lceil k/2 \rceil} e}{e'(A'A)^k A'e}. \quad (2a)$$

Observe that $2\lfloor k/2 \rfloor + 1 = k$ and $2\lceil k/2 \rceil = k + 1$ because k is odd. Also, $\alpha_{2k+1} = \beta_{2k+1}$. Thus, we can write equations (1a) and (2a) as

$$\frac{\beta_{2k+1}}{\beta_{2k}} \leq \frac{\alpha_{2k+2}}{\alpha_{2k+1}}, \quad (1b)$$

$$\frac{\alpha_{2k+1}}{\alpha_{2k}} \leq \frac{\beta_{2k+2}}{\beta_{2k+1}}. \quad (2b)$$

We use Lemma 1 to conclude that equation (1b) implies $E(d_{in}(X'_{2k})) \leq E(d_{in}(X_{2k+1}))$ and equation (2b) implies $E(d_{out}(X_{2k})) \leq E(d_{out}(X'_{2k+1}))$.

(2) If k is even, then we can rewrite equations (1) and (2) in the form

$$\frac{e'(AA')^k Ae}{e'(AA')^{2\lceil k/2 \rceil} e} \leq \frac{e'(A'A)^{2\lfloor k/2 \rfloor + 1} e}{e'(AA')^k Ae}, \quad (1a')$$

$$\frac{e'(A'A)^k A'e}{e'(A'A)^{2\lceil k/2 \rceil} e} \leq \frac{e'(AA')^{2\lfloor k/2 \rfloor + 1} e}{e'(A'A)^k A'e}. \quad (2a')$$

Observe that $2\lceil k/2 \rceil + 1 = k + 1$ and $2\lfloor k/2 \rfloor = k$ because k is even. In addition, $\alpha_{2k+1} = \beta_{2k+1}$. Thus, we can write equations (1a) and (2a) as

$$\frac{\alpha_{2k+1}}{\alpha_{2k}} \leq \frac{\beta_{2k+2}}{\beta_{2k+1}}, \quad (1b')$$

$$\frac{\beta_{2k+1}}{\beta_{2k}} \leq \frac{\alpha_{2k+2}}{\alpha_{2k+1}}. \quad (2b')$$

From Lemma 1, equation (1b') implies $E(d_{out}(X_{2k})) \leq E(d_{out}(X'_{2k+1}))$ and equation (2b') implies $E(d_{in}(X'_{2k})) \leq E(d_{in}(X_{2k+1}))$. \square

B2: Proof of Theorem 2

Let $A = UDV'$ be the singular value decomposition of A , where D is the diagonal matrix of singular values, and U and V are unitary matrices whose columns are the left and right singular vectors, respectively. Let σ_i denote the i th largest singular

value of A , v_i the i th column of V , and u_i the i th column of U , for all $i = 1, \dots, n$. Then we can express α_{2k} and β_{2k} as

$$\alpha_{2k} = (AA')^k = (UDV'VDU')^k = UD^{2k}U' = \sum_{i=1}^n \sigma_i^{2k} u_i u_i',$$

$$\beta_{2k} = (A'A)^k = (VDU'UDV')^k = VD^{2k}V' = \sum_{i=1}^n \sigma_i^{2k} v_i v_i'.$$

Similarly, we can express α_{2k+1} and β_{2k+1} as

$$\alpha_{2k+1} = (AA')^k A = UD^{2k}U'UDV' = UD^{2k+1}V' = \sum_{i=1}^n \sigma_i^{2k+1} u_i v_i',$$

$$\beta_{2k+1} = (A'A)^k A' = VD^{2k}V'VDU' = VD^{2k+1}U' = \sum_{i=1}^n \sigma_i^{2k+1} v_i u_i'.$$

From Lemma 1,

$$E(d_{out}(X_{2k})) = \frac{\alpha_{2k+1}}{\alpha_{2k}} = \frac{\sum_{i=1}^n \sigma_i^{2k+1} e' u_i v_i' e}{\sum_{i=1}^n \sigma_i^{2k} e' u_i u_i' e}.$$

Let $\sigma_1 = \sigma_2 = \dots = \sigma_l > \sigma_{l+1} \dots$. Then,

$$E(d_{out}(X_{2k})) = \sigma_1 \frac{\sum_{i=1}^l e' u_i v_i' e + \sum_{i=l+1}^n \left(\frac{\sigma_i}{\sigma_1}\right)^{2k+1} e' u_i v_i' e}{\sum_{i=1}^l e' u_i u_i' e + \sum_{i=l+1}^n \left(\frac{\sigma_i}{\sigma_1}\right)^{2k} e' u_i u_i' e}.$$

Since $\sigma_i/\sigma_1 < 1$ for all $i = l+1, \dots, n$,

$$\lim_{k \rightarrow \infty} E(d_{out}(X_{2k})) = \frac{\sum_{i=1}^l e' u_i v_i' e}{\sum_{i=1}^l e' u_i u_i' e} \sigma_1 = c_{out} \sigma_1.$$

A similar argument shows that

$$\lim_{k \rightarrow \infty} E(d_{out}(X'_{2k+1})) = \frac{\sum_{i=1}^l e' v_i v_i' e}{\sum_{i=1}^l e' v_i u_i' e} \sigma_1 = c'_{out} \sigma_1,$$

$$\lim_{k \rightarrow \infty} E(d_{in}(X'_{2k})) = \frac{\sum_{i=1}^l e' v_i u_i' e}{\sum_{i=1}^l e' v_i v_i' e} \sigma_1 = c'_{in} \sigma_1,$$

$$\lim_{k \rightarrow \infty} E(d_{in}(X_{2k+1})) = \frac{\sum_{i=1}^l e' u_i u_i' e}{\sum_{i=1}^l e' u_i v_i' e} \sigma_1 = c_{in} \sigma_1,$$

where $c_{out} = 1/c_{in}$ and $c'_{out} = 1/c'_{in}$. If σ_1 is non-degenerate, then

$$c_{out} = \frac{e'u_1v'_1e}{e'u_1u'_1e} = \frac{v'_1e}{u'_1e} = \frac{e'v_1v'_1e}{e'v_1u'_1e}\sigma_1 = c'_{out},$$

$$c_{in} = \frac{e'u_1u'_1e}{e'u_1v'_1e} = \frac{u'_1e}{v'_1e} = \frac{e'v_1u'_1e}{e'v_1v'_1e}\sigma_1 = c'_{in},$$

where $e'u_1 = u'_1e$ and $e'v_1 = v'_1e$ are scalars. Thus,

$$\lim_{k \rightarrow \infty} E(d_{out}(X_{2k})) = \lim_{k \rightarrow \infty} E(d_{out}(X'_{2k+1})) = c_{out}\sigma_1,$$

and

$$\lim_{k \rightarrow \infty} E(d_{in}(X'_{2k})) = \lim_{k \rightarrow \infty} E(d_{in}(X_{2k+1})) = c_{in}\sigma_1,$$

where $c_{out} = 1/c_{in}$. \square

Appendix C: Essay 3

C1: VAE Architecture

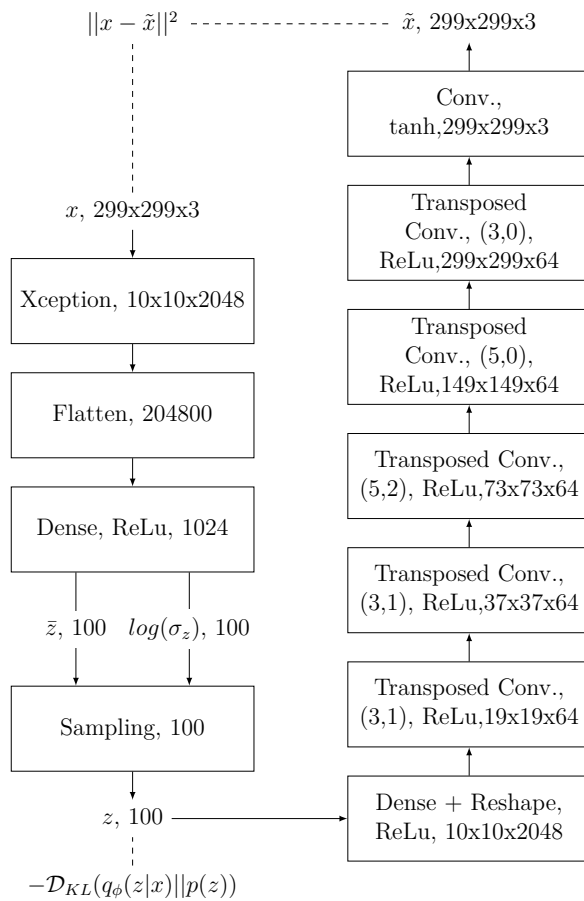


Figure 1: Variational Autoencoder Architecture. Transposed convolution layers have a kernel with 64 filters and a stride of size 2; remaining parameters: (kernel size, padding).

C2: Full Regression Results

C2a: Regular Auctions

Table 6: Regression Results for Regular Auctions

	<i>Dependent variable:</i>		
	$\ln(P)$		
	Auctions + Paintings	Baseline + Visual	Full Model
Intercept	17.789*** (0.161)	17.912*** (0.165)	18.353*** (0.177)
Artist dummies	[Included]	[Included]	[Included]
Authenticity dummies	[Included]	[Included]	[Included]
Bonhams	-1.441*** (0.036)	-1.411*** (0.036)	-1.415*** (0.036)
Christie's	-0.164*** (0.029)	-0.125*** (0.029)	-0.129*** (0.029)
Doyle	-1.990*** (0.033)	-1.940*** (0.034)	-1.947*** (0.034)
Gene Shapiro	-1.483*** (0.047)	-1.422*** (0.047)	-1.431*** (0.047)
Heritage Auctions	-1.638*** (0.050)	-1.665*** (0.050)	-1.690*** (0.050)
Sotheby's	0.181*** (0.029)	0.267*** (0.030)	0.271*** (0.030)
Swann	-0.976*** (0.048)	-0.977*** (0.048)	-0.988*** (0.048)
European	-0.037** (0.017)	-0.031* (0.017)	-0.028 (0.017)
American	0.203*** (0.016)	0.204*** (0.016)	0.206*** (0.016)
Old Masters	0.391*** (0.047)	0.368*** (0.047)	0.400*** (0.047)
Contemporary	0.172*** (0.021)	0.145*** (0.021)	0.139*** (0.021)
Modern	0.156*** (0.029)	0.158*** (0.029)	0.157*** (0.029)
PostWar	0.760*** (0.037)	0.763*** (0.037)	0.769*** (0.037)
Impressionist	-0.135*** (0.038)	-0.178*** (0.038)	-0.182*** (0.038)

Latin	-0.198*** (0.026)	-0.212*** (0.026)	-0.221*** (0.026)
Dogs	0.510*** (0.054)	0.515*** (0.054)	0.499*** (0.054)
Area	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
Date	-0.004*** (0.0001)	-0.005*** (0.0001)	-0.005*** (0.0001)
After	-1.823*** (0.065)	-1.819*** (0.065)	-1.809*** (0.065)
Attributed	-1.032*** (0.040)	-1.027*** (0.040)	-1.009*** (0.040)
Circle	-1.519*** (0.047)	-1.506*** (0.047)	-1.495*** (0.046)
Follower	-1.806*** (0.054)	-1.788*** (0.054)	-1.784*** (0.054)
Manner	-1.903*** (0.076)	-1.882*** (0.076)	-1.877*** (0.076)
School	-1.180*** (0.057)	-1.183*** (0.057)	-1.181*** (0.056)
Studio	-0.415*** (0.073)	-0.438*** (0.072)	-0.447*** (0.072)
Acrylic	-0.232*** (0.030)	-0.233*** (0.030)	-0.233*** (0.030)
Oil	-0.236*** (0.025)	-0.217*** (0.025)	-0.215*** (0.025)
Watercolor	-0.654*** (0.216)	-0.631*** (0.215)	-0.616*** (0.215)
Board	-0.074*** (0.022)	-0.077*** (0.022)	-0.074*** (0.022)
Canvas	0.331*** (0.014)	0.326*** (0.014)	0.329*** (0.014)
Wood	0.016 (0.049)	0.046 (0.048)	0.054 (0.048)
February	-0.468*** (0.029)	-0.436*** (0.029)	-0.440*** (0.029)
March	-0.057** (0.027)	-0.058** (0.027)	-0.061** (0.027)
April	0.253*** (0.026)	0.260*** (0.026)	0.257*** (0.026)
May	0.971*** (0.025)	0.964*** (0.025)	0.959*** (0.025)
June	-0.582*** (0.026)	-0.561*** (0.026)	-0.556*** (0.026)
July	-0.919***	-0.889***	-0.889***

	(0.043)	(0.043)	(0.043)
August	-1.440***	-1.343***	-1.340***
	(0.062)	(0.062)	(0.061)
September	-0.075***	-0.080***	-0.082***
	(0.026)	(0.026)	(0.026)
October	-0.082***	-0.073***	-0.076***
	(0.026)	(0.025)	(0.025)
November	0.934***	0.916***	0.912***
	(0.025)	(0.025)	(0.025)
December	0.161***	0.179***	0.177***
	(0.029)	(0.029)	(0.029)
2000	0.765***	0.797***	0.787***
	(0.079)	(0.079)	(0.078)
2001	0.422***	0.427***	0.420***
	(0.072)	(0.072)	(0.072)
2002	0.781***	0.794***	0.798***
	(0.071)	(0.071)	(0.071)
2003	0.580***	0.591***	0.591***
	(0.072)	(0.072)	(0.072)
2004	0.574***	0.586***	0.588***
	(0.070)	(0.070)	(0.070)
2005	0.697***	0.686***	0.691***
	(0.071)	(0.071)	(0.071)
2006	0.856***	0.868***	0.874***
	(0.070)	(0.069)	(0.069)
2007	0.987***	1.000***	1.007***
	(0.070)	(0.070)	(0.069)
2008	0.822***	0.837***	0.845***
	(0.070)	(0.070)	(0.070)
2009	0.388***	0.402***	0.410***
	(0.071)	(0.071)	(0.071)
2010	0.414***	0.435***	0.447***
	(0.070)	(0.070)	(0.070)
2011	0.435***	0.477***	0.490***
	(0.070)	(0.070)	(0.070)
2012	0.447***	0.485***	0.492***
	(0.070)	(0.071)	(0.071)
2013	0.619***	0.617***	0.626***
	(0.071)	(0.071)	(0.071)
2014	0.569***	0.577***	0.583***
	(0.070)	(0.070)	(0.070)
2015	0.536***	0.554***	0.560***
	(0.070)	(0.070)	(0.070)
2016	0.448***	0.463***	0.470***
	(0.071)	(0.071)	(0.071)

2017	0.569*** (0.072)	0.579*** (0.072)	0.589*** (0.072)
2018	0.600*** (0.072)	0.587*** (0.072)	0.595*** (0.072)
Visual Factor 1		-0.009*** (0.002)	-0.005** (0.002)
Visual Factor 2		0.024*** (0.002)	0.022*** (0.003)
Visual Factor 3		0.006** (0.003)	0.016*** (0.003)
Visual Factor 4		-0.041*** (0.003)	-0.042*** (0.003)
Visual Factor 5		-0.0005 (0.003)	-0.005 (0.003)
Visual Factor 6		0.019*** (0.003)	0.021*** (0.003)
Visual Factor 7		0.028*** (0.003)	0.026*** (0.003)
Visual Factor 8		0.033*** (0.003)	0.038*** (0.003)
Visual Factor 9		0.019*** (0.003)	0.018*** (0.003)
Visual Factor 10		0.006* (0.003)	0.007** (0.003)
Visual Factor 11		-0.002 (0.004)	-0.003 (0.004)
Visual Factor 12		-0.015*** (0.004)	-0.012*** (0.004)
Visual Factor 13		0.031*** (0.004)	0.028*** (0.004)
Visual Factor 14		0.001 (0.004)	-0.001 (0.004)
Visual Factor 15		-0.013*** (0.004)	-0.012*** (0.004)
Visual Factor 16		-0.001 (0.004)	0.001 (0.004)
Visual Factor 17		-0.004 (0.004)	-0.002 (0.004)
Visual Factor 18		0.005 (0.004)	0.006 (0.004)
Visual Factor 19		0.023*** (0.004)	0.022*** (0.004)
Visual Factor 20		-0.009** (0.004)	-0.011** (0.004)
Visual Factor 21		0.008* (0.004)	0.007* (0.004)

		(0.004)	(0.004)
Visual Factor 22		-0.006	-0.007
		(0.004)	(0.004)
Visual Factor 23		0.019***	0.020***
		(0.005)	(0.005)
Visual Factor 24		-0.016***	-0.015***
		(0.005)	(0.005)
Visual Factor 25		-0.009*	-0.009**
		(0.005)	(0.005)
Visual Factor 26		-0.013***	-0.013***
		(0.005)	(0.005)
Visual Factor 27		0.010**	0.005
		(0.005)	(0.005)
Visual Factor 28		0.014***	0.011**
		(0.005)	(0.005)
Visual Factor 29		0.028***	0.028***
		(0.005)	(0.005)
Visual Factor 30		0.003	0.003
		(0.005)	(0.005)
d_{out}^{other}			-0.0001***
			(0.00001)
d_{in}^{other}			-0.00001**
			(0.00000)
d_{out}^{same}			0.008***
			(0.003)
d_{in}^{same}			0.015***
			(0.003)
$d_{out/in}$			-0.00002**
			(0.00001)
$d_{in/out}$			0.00001**
			(0.00000)
Observations	71,215	71,215	71,215
R ²	0.492	0.498	0.499
Adjusted R ²	0.490	0.496	0.497
Residual Std. Error	1.322 (df = 70954)	1.314 (df = 70924)	1.312 (df = 70918)
F-Statistic	264.451***	242.521***	239.038***
	(df = 260; 70954)	(df = 290; 70924)	(df = 296; 70918)
In-Sample MAE	184,158.69	183,768.83	183,686.81
Out-Sample MAE	172,654.34	172,010.58	171,947.34

Note:

*p < 0.1; **p < 0.05; ***p < 0.01

C2b: Day Auctions

Table 7: Regression Results for Day Auctions

	<i>Dependent variable:</i>		
	$\ln(P)$		
	Auctions + Paintings	Baseline + Visual	Full Model
Intercept	15.629*** (0.645)	16.099*** (0.650)	15.546*** (0.695)
Artist dummies	[Included]	[Included]	[Included]
Authenticity dummies	[Included]	[Included]	[Included]
Christie's	0.711*** (0.049)	0.743*** (0.050)	0.728*** (0.050)
Doyle	-1.148*** (0.418)	-1.150*** (0.418)	-1.152*** (0.416)
Sotheby's	0.745*** (0.042)	0.739*** (0.042)	0.722*** (0.042)
European	-0.094 (0.162)	-0.088 (0.162)	-0.095 (0.161)
American	-0.840*** (0.185)	-0.825*** (0.185)	-0.822*** (0.184)
Contemporary	0.117 (0.114)	0.117 (0.114)	0.112 (0.114)
Modern	-0.396*** (0.114)	-0.352*** (0.114)	-0.364*** (0.114)
Post-War	0.086** (0.038)	0.059 (0.039)	0.053 (0.039)
Latin	-1.330*** (0.187)	-1.304*** (0.188)	-1.299*** (0.187)
Area	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
Date	-0.004*** (0.0004)	-0.004*** (0.0004)	-0.004*** (0.0004)
After	-0.826 (0.565)	-0.926 (0.564)	-0.942* (0.562)
Attributed	-0.686*** (0.263)	-0.729*** (0.262)	-0.747*** (0.262)
Circle	-0.794** (0.323)	-0.836*** (0.322)	-0.837*** (0.321)
Follower	-0.477 (0.442)	-0.547 (0.442)	-0.514 (0.441)
School	-0.305	-0.322	-0.323

	(0.307)	(0.307)	(0.306)
Studio	-1.130**	-1.193***	-1.139**
	(0.446)	(0.446)	(0.445)
Acrylic	-0.064	-0.061	-0.061
	(0.041)	(0.041)	(0.041)
Oil	0.107***	0.120***	0.116***
	(0.040)	(0.040)	(0.040)
Watercolor	0.262	0.238	0.275
	(0.350)	(0.350)	(0.349)
Board	0.153***	0.159***	0.139***
	(0.049)	(0.049)	(0.049)
Canvas	0.441***	0.442***	0.437***
	(0.028)	(0.028)	(0.028)
Wood	0.097	0.093	0.112
	(0.083)	(0.083)	(0.083)
February	0.271*	0.233	0.262*
	(0.148)	(0.149)	(0.148)
March	-1.055***	-1.088***	-1.080***
	(0.369)	(0.369)	(0.368)
April	-0.947***	-1.011***	-0.952***
	(0.275)	(0.276)	(0.275)
May	1.875***	1.828***	1.846***
	(0.175)	(0.176)	(0.175)
July	-2.770***	-2.750***	-2.734***
	(0.261)	(0.261)	(0.260)
September	-0.559	-0.580	-0.588
	(0.391)	(0.391)	(0.389)
October	0.218	0.196	0.201
	(0.202)	(0.202)	(0.202)
November	1.890***	1.846***	1.864***
	(0.176)	(0.176)	(0.176)
December	-0.763*	-0.743*	-0.746*
	(0.427)	(0.427)	(0.425)
2004	-0.268**	-0.271**	-0.281**
	(0.125)	(0.126)	(0.125)
2005	0.690***	0.706***	0.711***
	(0.118)	(0.118)	(0.117)
2006	0.699***	0.686***	0.659***
	(0.138)	(0.140)	(0.139)
2007	0.849***	0.850***	0.847***
	(0.097)	(0.097)	(0.097)
2008	0.743***	0.749***	0.736***
	(0.099)	(0.099)	(0.099)
2009	0.562***	0.574***	0.571***
	(0.099)	(0.099)	(0.098)

2010	0.590*** (0.098)	0.611*** (0.098)	0.610*** (0.098)
2011	0.582*** (0.099)	0.607*** (0.100)	0.606*** (0.099)
2012	0.583*** (0.096)	0.613*** (0.097)	0.614*** (0.096)
2013	0.656*** (0.096)	0.683*** (0.096)	0.676*** (0.095)
2014	0.720*** (0.095)	0.735*** (0.096)	0.728*** (0.095)
2015	0.702*** (0.096)	0.707*** (0.096)	0.704*** (0.096)
2016	0.603*** (0.097)	0.604*** (0.097)	0.600*** (0.097)
2017	0.613*** (0.096)	0.616*** (0.096)	0.615*** (0.096)
2018	0.750*** (0.096)	0.751*** (0.096)	0.740*** (0.096)
Visual Factor 1		0.016*** (0.006)	0.020*** (0.006)
Visual Factor 2		-0.018*** (0.006)	-0.017*** (0.006)
Visual Factor 3		-0.011** (0.005)	0.007 (0.006)
Visual Factor 4		-0.017*** (0.006)	-0.014** (0.006)
Visual Factor 5		0.015** (0.006)	0.011* (0.006)
Visual Factor 6		0.0004 (0.006)	0.002 (0.006)
Visual Factor 7		0.008 (0.006)	0.010* (0.006)
Visual Factor 8		0.012* (0.006)	0.019*** (0.006)
Visual Factor 9		0.008 (0.007)	0.007 (0.007)
Visual Factor 10		0.011 (0.007)	0.009 (0.007)
Visual Factor 11		-0.015* (0.008)	-0.013 (0.008)
Visual Factor 12		-0.016** (0.007)	-0.012 (0.007)
Visual Factor 13		0.015* (0.008)	0.011 (0.008)
Visual Factor 14		-0.016* (0.008)	-0.018** (0.008)

	(0.008)	(0.008)
Visual Factor 15	-0.007	-0.005
	(0.008)	(0.008)
Visual Factor 16	0.009	0.011
	(0.008)	(0.008)
Visual Factor 17	0.007	0.009
	(0.008)	(0.008)
Visual Factor 18	-0.004	-0.005
	(0.008)	(0.008)
Visual Factor 19	-0.001	-0.003
	(0.009)	(0.009)
Visual Factor 20	0.014	0.014
	(0.009)	(0.009)
Visual Factor 21	-0.011	-0.012
	(0.009)	(0.009)
Visual Factor 22	0.009	0.007
	(0.009)	(0.009)
Visual Factor 23	0.010	0.010
	(0.009)	(0.009)
Visual Factor 24	-0.005	-0.002
	(0.008)	(0.008)
Visual Factor 25	-0.019**	-0.020**
	(0.009)	(0.009)
Visual Factor 26	-0.008	-0.008
	(0.009)	(0.009)
Visual Factor 27	0.015	0.011
	(0.009)	(0.009)
Visual Factor 28	-0.012	-0.014
	(0.009)	(0.009)
Visual Factor 29	0.006	0.007
	(0.009)	(0.009)
Visual Factor 30	0.006	0.005
	(0.009)	(0.009)
d_{out}^{other}		0.00002
		(0.00002)
d_{in}^{other}		-0.00004***
		(0.00001)
d_{out}^{same}		0.007*
		(0.004)
d_{in}^{same}		0.026***
		(0.004)
$d_{out/in}$		0.00000
		(0.00001)
$d_{in/out}$		-0.00001*

			(0.00001)
Observations	10,545	10,545	10,545
R ²	0.461	0.465	0.470
Adjusted R ²	0.451	0.453	0.458
Residual Std. Error	0.967 (df = 10347)	0.965 (df = 10317)	0.961 (df = 10311)
F-Statistic	44.923*** (df = 197; 10347)	39.522*** (df = 227; 10317)	39.188*** (df = 233; 10311)
In-Sample MAE	187,707.32	187,171.54	186,454.67
Out-Sample MAE	167,375.56	167,125.84	166,567.33

Note:

*p < 0.1; **p < 0.05; ***p < 0.01

C2c: Evening Auctions

Table 8: Regression Results for Evening Auctions

	<i>Dependent variable:</i>		
	$\ln(P)$		
	Auctions + Paintings	Baseline + Visual	Full Model
Intercept	18.038*** (1.137)	18.823*** (1.160)	18.389*** (1.183)
Artist dummies	[Included]	[Included]	[Included]
Authenticity dummies	[Included]	[Included]	[Included]
Christie's	0.554*** (0.092)	0.540*** (0.093)	0.523*** (0.093)
Sotheby's	0.662*** (0.075)	0.629*** (0.077)	0.612*** (0.077)
American	-0.129 (0.218)	-0.220 (0.219)	-0.244 (0.219)
Contemporary	-0.018 (0.121)	-0.029 (0.122)	-0.022 (0.122)
Modern	0.095 (0.372)	0.121 (0.372)	0.141 (0.371)
PostWar	0.195** (0.078)	0.138* (0.079)	0.126 (0.079)
Impressionist	-0.262 (0.389)	-0.246 (0.389)	-0.265 (0.388)
Latin	-2.762*** (0.205)	-2.671*** (0.207)	-2.666*** (0.206)
Area	-0.000 (0.00000)	0.000 (0.00000)	0.000 (0.00000)
Date	-0.005*** (0.001)	-0.005*** (0.001)	-0.005*** (0.001)
Attributed	-0.620 (0.618)	-0.863 (0.616)	-0.784 (0.618)
Circle	0.408 (1.045)	0.469 (1.043)	0.396 (1.042)
Follower	-0.597 (0.534)	-0.624 (0.532)	-0.565 (0.533)
School	-0.512 (0.443)	-0.436 (0.442)	-0.435 (0.441)
Acrylic	-0.112 (0.086)	-0.099 (0.086)	-0.143 (0.088)
Oil	0.199**	0.230***	0.200**

	(0.078)	(0.078)	(0.079)
Watercolor	-0.431	-0.450	-0.405
	(1.072)	(1.069)	(1.068)
Board	-0.045	-0.046	-0.054
	(0.118)	(0.118)	(0.118)
Canvas	0.427***	0.421***	0.419***
	(0.064)	(0.064)	(0.064)
Wood	0.131	0.092	0.099
	(0.181)	(0.180)	(0.180)
February	0.101	0.004	0.030
	(0.237)	(0.238)	(0.238)
March	1.571***	1.608***	1.597***
	(0.348)	(0.349)	(0.349)
May	3.697***	3.668***	3.633***
	(0.302)	(0.303)	(0.304)
October	-0.811	-0.661	-0.592
	(1.113)	(1.110)	(1.108)
November	3.613***	3.593***	3.559***
	(0.303)	(0.304)	(0.304)
December	-0.380	-0.721	-0.766
	(1.068)	(1.070)	(1.068)
2002	0.759***	0.702***	0.744***
	(0.239)	(0.240)	(0.239)
2003	0.536**	0.491**	0.522**
	(0.237)	(0.238)	(0.237)
2004	1.102***	1.056***	1.109***
	(0.282)	(0.283)	(0.283)
2005	1.411***	1.458***	1.468***
	(0.260)	(0.260)	(0.260)
2006	1.188***	1.111***	1.137***
	(0.259)	(0.262)	(0.261)
2007	2.051***	1.959***	2.001***
	(0.202)	(0.205)	(0.204)
2008	1.342***	1.310***	1.347***
	(0.196)	(0.197)	(0.197)
2009	1.482***	1.406***	1.438***
	(0.197)	(0.199)	(0.199)
2010	1.618***	1.577***	1.611***
	(0.193)	(0.194)	(0.194)
2011	1.685***	1.633***	1.667***
	(0.191)	(0.193)	(0.193)
2012	1.940***	1.880***	1.907***
	(0.191)	(0.193)	(0.192)
2013	1.908***	1.861***	1.888***
	(0.187)	(0.189)	(0.189)

2014	2.094*** (0.188)	2.044*** (0.190)	2.073*** (0.190)
2015	2.108*** (0.187)	2.081*** (0.190)	2.109*** (0.189)
2016	2.049*** (0.190)	1.996*** (0.193)	2.025*** (0.193)
2017	2.137*** (0.189)	2.106*** (0.191)	2.133*** (0.191)
2018	2.093*** (0.184)	2.065*** (0.187)	2.089*** (0.187)
Visual Factor 1		0.034*** (0.012)	0.037*** (0.012)
Visual Factor 2		-0.032** (0.013)	-0.030** (0.013)
Visual Factor 3		-0.017 (0.010)	0.005 (0.012)
Visual Factor 4		-0.018 (0.012)	-0.015 (0.012)
Visual Factor 5		-0.002 (0.012)	-0.006 (0.012)
Visual Factor 6		0.026* (0.014)	0.030** (0.014)
Visual Factor 7		0.012 (0.012)	0.015 (0.012)
Visual Factor 8		0.004 (0.013)	0.015 (0.013)
Visual Factor 9		-0.019 (0.014)	-0.020 (0.014)
Visual Factor 10		0.041*** (0.014)	0.041*** (0.014)
Visual Factor 11		-0.046*** (0.015)	-0.045*** (0.015)
Visual Factor 12		0.014 (0.014)	0.020 (0.014)
Visual Factor 13		0.041*** (0.015)	0.037** (0.015)
Visual Factor 14		0.001 (0.016)	-0.001 (0.016)
Visual Factor 15		-0.028* (0.015)	-0.023 (0.015)
Visual Factor 16		-0.0002 (0.016)	0.003 (0.016)
Visual Factor 17		0.008 (0.017)	0.009 (0.017)
Visual Factor 18		-0.001	-0.004

		(0.017)	(0.017)
Visual Factor 19		-0.019	-0.020
		(0.017)	(0.017)
Visual Factor 20		0.019	0.016
		(0.017)	(0.017)
Visual Factor 21		-0.043**	-0.041**
		(0.017)	(0.017)
Visual Factor 22		0.008	0.006
		(0.017)	(0.017)
Visual Factor 23		0.018	0.020
		(0.017)	(0.017)
Visual Factor 24		-0.021	-0.019
		(0.016)	(0.016)
Visual Factor 25		0.008	0.006
		(0.018)	(0.018)
Visual Factor 26		0.018	0.017
		(0.018)	(0.018)
Visual Factor 27		-0.013	-0.018
		(0.018)	(0.018)
Visual Factor 28		0.006	0.001
		(0.019)	(0.019)
Visual Factor 29		0.019	0.020
		(0.019)	(0.019)
Visual Factor 30		-0.031*	-0.033*
		(0.019)	(0.019)
d_{out}^{other}			0.00001
			(0.00003)
d_{in}^{other}			-0.00004***
			(0.00002)
d_{out}^{same}			0.003
			(0.005)
d_{in}^{same}			0.011**
			(0.004)
$d_{out/in}$			-0.00001
			(0.00003)
$d_{in/out}$			-0.00002
			(0.00002)
Observations	3,152	3,152	3,152
R ²	0.472	0.484	0.488
Adjusted R ²	0.448	0.455	0.458
Residual Std. Error	1.029 (df = 3011)	1.023 (df = 2981)	1.020 (df = 2975)
F-Statistic	19.256***)	16.447***	16.098***
	(df = 140; 3011)	(df = 170; 2981)	(df = 176; 2975)
In-Sample MAE	4,395,924.33	4,364,960.39	4,354,208.72

Out-Sample MAE	5,235,104.11	5,194,797.06	5,181,380.11
----------------	--------------	--------------	--------------

Note:

*p < 0.1; **p < 0.05; ***p < 0.01

C3: Matching Performance

Table 9: Mean Comparisons between Control and Treatment Group after Matching

Variables	Control Group (n=3966)	Treatment Group (n=3966)	P-Value
	Mean (S.D.)	Mean (S.D.)	
Basquiat Jean-Michel	0.01 (0.09)	0.02 (0.15)	0
Lichtenstein Roy	0.01 (0.1)	0.02 (0.14)	0
Matisse Henri	0.01 (0.08)	0.01 (0.1)	0.02
Miró Joan	0.01 (0.08)	0.01 (0.11)	0.01
Monet Claude	0 (0.06)	0.03 (0.17)	0
Picasso Pablo	0.01 (0.11)	0.05 (0.21)	0
Pissarro Camille	0.01 (0.1)	0.02 (0.13)	0.02
Richter Gerhard	0.02 (0.14)	0.03 (0.16)	0.01
Rothko Mark	0 (0.04)	0.01 (0.12)	0
Sisley Alfred	0.01 (0.08)	0.01 (0.11)	0.01
Warhol Andy	0.08 (0.28)	0.06 (0.24)	0
Remaining Artists (78)	—	—	≥ 0.05
Initials	0.03 (0.18)	0.03 (0.16)	0.12
Inscription	0 (0.05)	0 (0.05)	0.65
Month	0.01 (0.11)	0.01 (0.1)	0.29
Signature	0.79 (0.41)	0.79 (0.41)	0.89
Christies	0.38 (0.49)	0.44 (0.5)	0
Sothebys	0.49 (0.5)	0.45 (0.5)	0
American	0.02 (0.14)	0.02 (0.14)	1
Contemporary	0.6 (0.49)	0.57 (0.5)	0
Modern	0.33 (0.47)	0.35 (0.48)	0.14
Post War	0.22 (0.41)	0.23 (0.42)	0.19
Latin	0.02 (0.14)	0.02 (0.14)	0.94
Area	23319.75 (201603)	25746.23 (195492.19)	0.59
Date	1950.05 (70.98)	1946.42 (70.43)	0.02
Attributed	0 (0.04)	0 (0.03)	0.74
Circle	0 (0.02)	0 (0.02)	1
Follower	0 (0.04)	0 (0.03)	0.74
School	0 (0.05)	0 (0.04)	0.81
Studio	0 (0.02)	0 (0.02)	0.56
Acrylic	0.15 (0.35)	0.13 (0.34)	0.1
Oil	0.73 (0.45)	0.76 (0.43)	0
Watercolor	0 (0.02)	0 (0.02)	1
Board	0.04 (0.2)	0.04 (0.19)	0.27
Canvas	0.82 (0.39)	0.83 (0.37)	0.12
Wood	0.01 (0.11)	0.01 (0.11)	0.92
February	0.01 (0.12)	0.02 (0.12)	0.71

March	0 (0.05)	0.01 (0.11)	0
May	0.44 (0.5)	0.43 (0.49)	0.36
October	0 (0.02)	0 (0.02)	0.56
November	0.53 (0.5)	0.53 (0.5)	0.98
December	0 (0.02)	0 (0.02)	1
2002	0 (0)	0.02 (0.12)	0
2003	0.01 (0.12)	0.01 (0.12)	0.92
2004	0.01 (0.1)	0.01 (0.08)	0.17
2005	0.01 (0.1)	0.01 (0.09)	0.36
2006	0.01 (0.1)	0.01 (0.1)	0.91
2007	0.04 (0.2)	0.04 (0.21)	0.37
2008	0.07 (0.26)	0.06 (0.24)	0.09
2009	0.05 (0.22)	0.05 (0.22)	0.58
2010	0.06 (0.23)	0.06 (0.25)	0.12
2011	0.07 (0.25)	0.07 (0.25)	0.53
2012	0.07 (0.25)	0.07 (0.25)	0.62
2013	0.09 (0.29)	0.09 (0.28)	0.58
2014	0.09 (0.29)	0.09 (0.29)	0.91
2015	0.09 (0.29)	0.09 (0.28)	0.27
2016	0.09 (0.28)	0.09 (0.28)	0.94
2017	0.12 (0.33)	0.11 (0.31)	0.12
2018	0.11 (0.32)	0.11 (0.31)	0.35
Visual Factor 1	0.45 (1.98)	0.41 (1.83)	0.42
Visual Factor 2	-0.33 (1.86)	-0.3 (1.71)	0.58
Visual Factor 3	-0.66 (2.06)	-0.67 (2.1)	0.8
Visual Factor 4	-0.29 (1.76)	-0.3 (1.7)	0.85
Visual Factor 5	0.54 (1.75)	0.59 (1.82)	0.3
Visual Factor 6	0.07 (1.54)	0.13 (1.47)	0.08
Visual Factor 7	0.18 (1.67)	0.22 (1.7)	0.3
Visual Factor 8	0.05 (1.56)	0.08 (1.55)	0.39
Visual Factor 9	0.16 (1.42)	0.14 (1.39)	0.54
Visual Factor 10	0.16 (1.45)	0.23 (1.43)	0.04
Visual Factor 11	-0.08 (1.31)	-0.12 (1.35)	0.12
Visual Factor 12	-0.12 (1.52)	-0.07 (1.51)	0.13
Visual Factor 13	0.22 (1.24)	0.29 (1.31)	0.01
Visual Factor 14	0 (1.21)	-0.02 (1.21)	0.32
Visual Factor 15	-0.11 (1.25)	-0.13 (1.3)	0.49
Visual Factor 16	0.02 (1.17)	0.01 (1.18)	0.91
Visual Factor 17	-0.02 (1.1)	0.01 (1.1)	0.28
Visual Factor 18	-0.03 (1.14)	-0.03 (1.16)	0.97
Visual Factor 19	-0.06 (1.15)	-0.08 (1.14)	0.56
Visual Factor 20	0 (1.14)	-0.01 (1.14)	0.76
Visual Factor 21	-0.23 (1.1)	-0.23 (1.14)	0.93
Visual Factor 22	-0.12 (1.11)	-0.13 (1.1)	0.63
Visual Factor 23	0.1 (1.09)	0.09 (1.13)	0.85

Visual Factor 24	-0.06 (1.16)	0 (1.2)	0.05
Visual Factor 25	-0.01 (1.08)	-0.02 (1.07)	0.62
Visual Factor 26	0 (1.04)	0 (1.02)	0.91
Visual Factor 27	0.14 (1.03)	0.15 (1.04)	0.75
Visual Factor 28	-0.03 (1.04)	-0.05 (1.02)	0.43
Visual Factor 29	0.04 (1.04)	0.03 (1.02)	0.79
Visual Factor 30	-0.04 (1.03)	-0.04 (1.01)	0.78
d_{out}^{other}	433.95 (904.87)	445.72 (929.57)	0.57
d_{in}^{other}	679.05 (1471.09)	626.69 (1549.89)	0.12
d_{out}^{same}	0.95 (4.17)	1.08 (5.17)	0.22
d_{in}^{same}	0.98 (4.52)	1.04 (5.38)	0.57
$d_{out/in}$	1486.57 (1503.09)	1444.28 (1513.28)	0.21
$d_{in/out}$	2116.27 (2231.85)	2070.15 (2251.92)	0.36
