

# **Network Structure and its Influence on User Behavior**

Brendan Meeder

CMU-CS-15-106

May 4, 2015

School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213

## **Thesis Committee:**

Luis von Ahn, co-chair

Manuel Blum, co-chair

Christos Faloutsos

Jon Kleinberg (Cornell University)

*Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy.*

Copyright © 2015 Brendan Meeder

This material is based upon the work supported by the National Science Foundation Graduate Research Fellowship Program under grant number DGE-0750271; National Science Foundation under grant numbers IIS-0082339, ISS-0222875, IIS-0910453; the Fine Foundation under grant number 083R17SM; and the Alfred P. Sloan Foundation under grant number BR4943. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

**Keywords:** Social networks, network analysis, graph mining, Duolingo, online education, language learning

*To my family, immediate and extended, for your never-ending support and encouragement.*



## **Abstract**

Social networks are now ubiquitous across many online services in diverse areas such as communication, music, education, health, and news. This thesis provides an understanding of the interplay between network structure and user behavior in the Twitter network and on the Duolingo language learning platform. We develop models and algorithms that can explain observed user behavior and distinguish between different patterns of behavior.

In the context of Twitter we present a model for diffusion that captures differences across topics such as news, politics, and popular culture and examine the initial conditions required to produce large information cascades. Additionally, we devise an algorithm for accurately inferring the times at which edges are created in a large subgraph of the social network. Using these inferred edge creation times we evaluate several models of network formation, see the impact of real-world events such as news stories on network evolution, and show how user interface design choices such as ‘Suggested Followers’ lists impact users’ following behavior.

Duolingo is the world’s largest language learning service and provides a unique opportunity to understand the impact of social features in an online education platform. Features such as an activity stream, leaderboard, and community sections exist to promote student-student interactions and keep students engaged in the learning process. We study the impact of these features on student behaviors and engagement over long periods of time (greater than six months). The techniques developed in this thesis can be used by creators of social systems to measure, model, and understand the impact of their design decisions.



## Acknowledgments

My time at Carnegie Mellon has spanned nearly fifteen years during which many people impacted my growth as a computer scientist. Matt Mason and Steven Rudich started the fire in me that became my passion for computer science with the Andrews' Leap program. Working on research projects with Matt Mason, Klaus Sutner and Manuel Blum solidified my intentions to enter a PhD program. Luis von Ahn and Klaus Sutner provided me the opportunity to take on increasing amounts of teaching responsibility by TAing their courses, giving lectures, and serving as a full instructor. Of all of my experiences at CMU I enjoyed serving as a mentor to hundreds of students most. I want to call out Mark Stehlik for being an excellent undergraduate advisor and a close friend over the last decade.

My graduate experience took an unexpected turn when almost three years into the program Luis invited me to build Duolingo with him and Severin Hacker. For the last four years Duolingo has been my obsession and an experience that helped me grow as a researcher and engineer. To my advisors Luis and Manuel: thank you for your inspiration and guidance. You both have a really hands-off style, but always helped me build connections with members of the research community through collaborations and internships. Having both of you in my academic and professional life for over a decade has been an honor. I am fortunate to have an outstanding set of peer collaborators: Daniel Romero, Brian Karrer, Amin Sayedi, Leman Akoglu and U Kang. To those established in academia, thank you for your mentorship and research ideas: Jon Kleinberg, Christos Faloutsos, R Ravi, Jennifer Chayes and Christian Borgs. I'm grateful to my world-class thesis committee for their insights, encouragement, and patience as I worked to finish this endeavor.

I give thanks to the entire Duolingo team for the dedication each person has had to making the best language learning experience available to everyone. Never have I been able to impact so many people, in such a meaningful way, as I have through my work at Duolingo. I'd like to particularly thank Severin Hacker, Matt Streeter, Vicki Cheung, and Tony Navas for helping me grow as an engineer, as well as Matt Streeter and Burr Settles for helping me pursue my research interests at Duolingo. Marcel, Myra, Pam, Monica, Kristine, and Max: thank you for regularly helping me reach my daily caffeine requirements and for the laughs we shared together. To all of my friends in the PhD program: thank you for being constantly in my life through the ups and downs of graduate school. I give a special thanks to Ekaterina Taralova, Matt Stanton, and Erik Zawadzki for their companionship. Finally, I am grateful to my wife Ariel Levavi for supporting me in everything I do in life and being my foundation. I wouldn't have been able to finish this journey without your love, encouragement and support.





# Contents

- 1 Introduction 1**
  - 1.1 Motivation . . . . . 1
  - 1.2 Contributions . . . . . 2
  
- 2 Network Structure and the Spread of Hashtags 5**
  - 2.1 Background . . . . . 5
  - 2.2 Dataset, Network Definition, and Hashtag Classification . . . . . 11
  - 2.3 Exposure Curves . . . . . 13
  - 2.4 The structure of initial sets . . . . . 19
  - 2.5 Simulations . . . . . 21
    - 2.5.1 The Simulated Model . . . . . 22
    - 2.5.2 Simulation Results . . . . . 22
  
- 3 The Evolution of Twitter’s Social Graph 25**
  - 3.1 Background . . . . . 25
  - 3.2 The Algorithm . . . . . 28
    - 3.2.1 Theoretical analysis . . . . . 29
  - 3.3 Application of the Algorithm . . . . . 30
    - 3.3.1 Broad analysis of celebrity subgraph . . . . . 30
    - 3.3.2 Impact of the Suggested Users List . . . . . 33
    - 3.3.3 Measuring following latency . . . . . 35
    - 3.3.4 Celebrity popularity and real-world events . . . . . 37
  - 3.4 Proof of proposition . . . . . 40
  
- 4 User Interactions in Twitter 43**
  - 4.1 Introduction . . . . . 43
    - 4.1.1 Summary of Results . . . . . 44
    - 4.1.2 Related work . . . . . 45
  - 4.2 Problem definition . . . . . 45
    - 4.2.1 Notation . . . . . 46
    - 4.2.2 Dataset description . . . . . 46
  - 4.3 Methods for Reciprocity Prediction . . . . . 47
    - 4.3.1 Degree and message features . . . . . 48
    - 4.3.2 Link prediction features . . . . . 49

4.3.3	Different sets of features . . . . .	50
4.3.4	Two-step paths . . . . .	50
4.4	Results and Discussion . . . . .	51
4.4.1	Individual properties . . . . .	51
4.4.2	Decision tree analysis . . . . .	54
4.4.3	Regression analysis . . . . .	57
4.5	Twitter as a superposition of networks . . . . .	58
4.5.1	(Un)reciprocated subgraph analysis . . . . .	58
<b>5</b>	<b>Duolingo: Massive Online Language Education</b>	<b>63</b>
5.1	Platform Overview . . . . .	63
5.1.1	Teaching Objectives . . . . .	64
5.1.2	Student Learning History . . . . .	64
5.1.3	Personalized Lessons . . . . .	65
5.1.4	Intelligent Feedback . . . . .	66
5.2	Game Mechanics . . . . .	67
5.2.1	eXperience Points and Levels . . . . .	67
5.2.2	The Coach and Streaks . . . . .	67
5.2.3	Lingots . . . . .	68
5.3	Social Features in Duolingo . . . . .	68
5.3.1	Finding Other Students on Duolingo . . . . .	70
5.4	Community Features in Duolingo . . . . .	71
5.5	Measurement and Experimentation . . . . .	72
5.6	Infrastructure . . . . .	74
<b>6</b>	<b>Duolingo's Social Network</b>	<b>77</b>
6.1	Breaking down the Student Body . . . . .	77
6.1.1	Languages learned . . . . .	78
6.2	Network Structure . . . . .	78
6.2.1	Degree distribution, connected components . . . . .	79
6.3	Network Evolution . . . . .	81
6.3.1	Densification . . . . .	81
6.3.2	Edge creation delay . . . . .	83
6.3.3	Other Following Behaviors . . . . .	87
<b>7</b>	<b>Student Behavior in Duolingo</b>	<b>89</b>
7.1	The Impact of Using Social Features . . . . .	89
7.1.1	Impact on Progress . . . . .	90
7.1.2	Impact on Student Lifetime . . . . .	93
7.2	Streaks and Continued Use . . . . .	96
7.2.1	Streak Length and Platform Usage . . . . .	96

<b>8 Conclusion</b>	<b>101</b>
8.1 Topical Differences in Information Diffusion . . . . .	101
8.2 Network Evolution in a Large Subgraph . . . . .	102
8.3 Social Networks in Online Education Platforms . . . . .	103
8.4 Future work: Increased Understanding of Student Behavior . . . . .	103
8.5 A Few Afterthoughts . . . . .	104
<b>Bibliography</b>	<b>107</b>



# List of Figures

- 2.1 Average exposure curve for the top 500 hashtags.  $P(K)$  is the fraction of users who adopt the hashtag directly after their  $k^{th}$  exposure to it, given that they had not yet adopted it . . . . . 8
- 2.2  $F(P)$  for the different types of hashtags. The black dots are the average  $F(P)$  among all hashtags, the red x is the average for the specific category, and the green dots indicate the 90% expected interval where the average for the specific set of hashtags would be if the set was chosen at random. Each point is the average of a set of at least 10 hashtags . . . . . 14
- 2.3 Sample exposure curves for hashtags #cantlivewithout (blue) and #hcr (red). . . . 15
- 2.4 Point-wise average influence curves. The blue line is the average of all the influence curves, the red line is the average for the set of hashtags of the particular topic, and the green lines indicate the interval where the red line is expected to be if the hashtags were chosen at random. . . . . 16
- 2.5 Example of the approximation of an influence curve. The red curve is the influence curve for the hashtag #pickone, the green curves indicate the 95% binomial confidence interval, and the blue curve is the approximation. . . . . 18
- 2.6 **Validating Category Differences:** The median cascade sizes for three different categories. In 2.6a we show the difference between celebrity and random  $p(k)$  curves on celebrity starting sets. celebrity hashtags and random starting sets. In 2.6b we show politics. Finally, we show idioms in 2.6c. All starting sets consist of 500 users. . . . . 23
- 3.1 Left side: The complementary cumulative distribution function for the number of followers of a celebrity. Note that this is a log-linear scale. Right side: The distribution of the number of celebrities followed by a user plotted on a log-log scale. Notice the three peaks at  $k = 20, 241$  and  $461$ . . . . . 32
- 3.2 The total celebrity follow rate (follow events per hour) and Twitter account creation rate (accounts created per day) over time. The three labels correspond to the introduction of the suggested users list, the update to the suggested users list, and introduction of “users you may be interested in”. The black smoothed curve shows a four day average of the celebrity follow rate. . . . . 33
- 3.3 The fraction of follow events for each celebrity per day as a function of time. The three labeled grey lines are the times of the interface changes described in Sec. 3.3.1. . . . . 34

3.4	The number of follow events binned by hour as a function of latency for the follow events of users created before September 1, 2010. . . . .	36
3.5	A heatmap of the creation time versus follow time over all celebrities with latencies greater than one day on a log-scale. The hours represent the GMT timezone. . . . .	37
3.6	The relative popularity as a function of time for the top 10 celebrities. The random attachment prediction is shown in bold. Labeled arrows correspond to events discussed in the text. . . . .	39
4.1	Proportion of nodes or edges (varying $n$ ) . . . . .	60
4.2	Proportion of nodes or edges (varying $k$ ) . . . . .	60
4.3	Clustering coefficient . . . . .	61
4.4	Proportion in largest connected component . . . . .	61
4.5	Scatter plot of users' interaction types . . . . .	61
5.1	The start of the German skill tree. . . . .	65
5.2	Strengthening skills in practice. . . . .	66
5.3	Flair for a student in the discussion forums. . . . .	68
5.4	The activity stream in Duolingo. . . . .	69
5.5	The Duolingo leaderboard. . . . .	70
5.6	The notification area displays recent events. . . . .	70
5.7	Using Facebook to find friends on Duolingo. . . . .	71
5.8	The main discussion view shows posts from subscribed topics. . . . .	72
5.9	A comment thread showing flair and additional social features. . . . .	73
5.10	Simplified architecture diagram of Duolingo. . . . .	75
6.1	Degree distributions in Duolingo's social network. . . . .	80
6.2	Comparison of the best log-normal and power-law fits with $D_{min} = 10$ . . . . .	81
6.3	Comparing the number of edges vs. the number of nodes in the network for each day since December 1, 2011. . . . .	82
6.4	Average out-degree vs. number of nodes for each day since December 1, 2011. . . . .	83
6.5	The number of edges created $H$ hours after account creation. . . . .	84
6.6	The number of edges created $d$ days after account creation. . . . .	85
6.7	New edges vs. daily accounts created . . . . .	86
6.8	New edges vs. daily registrations . . . . .	86
6.9	Reciprocated edge latency. We plot the fraction of reciprocated edges that form within $L$ hours of the initial edge forming. . . . .	87
6.10	The hourly rate at which new edges are created during the suggested user experiment. . . . .	88
7.1	The empirical and fitted lognormal distribution for the number of lessons completed. . . . .	90
7.2	Histogram of number of lessons completed based on connecting one's account to Facebook. . . . .	91
7.3	Histogram of number of lessons completed based on following or being followed. . . . .	92
7.4	Histogram of number of lessons completed based on different following conditions. . . . .	92

7.5	Student lifetime based on Facebook connectivity. . . . .	93
7.6	Student lifetimes based on following or being followed. . . . .	94
7.7	Student lifetime for students following, with different time requirements for when the follow events occur. . . . .	95
7.8	Empirical continuation probabilities for users in the first ten days of March, 2015.	97
7.9	Distribution of streak lengths and the best-fit power-law distribution ( $\alpha = -1.7670$ , $D_{min} = 6$ ) . . . . .	98
7.10	Proportions of students per platform, bucketed by streak length in days. . . . .	99
7.11	Proportions of students per platform, bucketed by streak length in weeks. The dotted lines mark the multinomial 95th percentile ranges for each proportion. . .	99





# List of Tables

2.1	Definitions of categories used for annotation. . . . .	12
2.2	A small set of examples of members in each category. . . . .	12
2.3	Median values for number of mentions, number of users, and number of mentions per user for different types of hashtags . . . . .	19
2.4	Comparison of graphs induced by the first 500 early adopters of political hashtags and average hashtags. Column definitions: I. Average degree, II. Average triangle count, III. Average entering degree of the nodes in the border of the graphs, IV. Average number of nodes in the border of the graphs. The error bars indicate the 95% confidence interval of the average value of a randomly selected set of hashtags of the same size as Political. . . . .	21
4.1	Reciprocity Prediction Features . . . . .	52
4.2	Indegree performance - different methods . . . . .	54
4.3	Reciprocity Prediction Feature Performance: Individual (REV) . . . . .	55
4.4	Reciprocity Prediction Feature Performance: Individual (REV- $v$ ,REV- $w$ ) . . . . .	56
4.5	Decision Tree Accuracy . . . . .	56
4.6	Logistic regression – relative degree/message-based features . . . . .	57
4.7	Logistic regression – two-step hop features . . . . .	58
4.8	Logistic regression – All ratio . . . . .	58
4.9	Logistic regression - All . . . . .	62
6.1	Account statistics as of November 23, 2014. . . . .	77
6.2	The most popular languages learned on Duolingo between September 1, 2014 and October 30, 2014 . . . . .	78
6.3	Network statistics for the Duolingo network as of November 23, 2014. . . . .	79



# Chapter 1

## Introduction

### 1.1 Motivation

Online social networks are ubiquitous today in an ever increasing number of services and products. Traditional online social networks, such as Facebook and Orkut are examples of systems that connect hundreds of millions to billions of people. In these networks users who are connected often have some relationship in the offline world. Other networks, such as Twitter, Google+, Flickr (photo sharing), and Pinterest (creative idea sharing) are a hybrid between social and information networks. Although individuals who know each other in real life might follow each other on these services, many connections exist between strangers because of interest in the content one or both produces. Some products and services, such as MyFitnessPal (a nutrition and weight loss site) and Duolingo (language learning), are not inherently social yet have social features to keep users engaged and encourage them to reach their goals. In all of these cases the structure of the social network, as well as how the network is integrated into the product, influences user behavior. This thesis examines the interplay between network structure and user behavior in Twitter and in Duolingo.

Twitter is an excellent environment in which user behavior can be observed because it has both a network structure and clear user actions. The network structure is defined by the *follower* relationship; a user sees content from all of the users they follow. This content is in the form of short messages called *tweets* that are 140 characters or less. Despite being quite short, tweets contain a surprising amount of information and context. For example, users can directly reference each other by *@-mentioning* each other. This behavior is unambiguous, is well structured, and

is easy to extract from the message. Additional semantic features such as *hashtags* (tokens that start with a # symbol) and URLs are also units of information that carry specific meaning and are easily extracted. Finally, messages can be repeated by *retweeting*, thereby expanding the visibility of a tweet in the network. As accounts are public by default, and only a small fraction of users make their tweets private, we can have a nearly complete view of the network, its users, and content therein. These qualities of Twitter come together to make it possible to study in detail information cascades, user-user interactions, network evolution, and general user behavior.

Duolingo is the world’s largest language learning platform, reaching over fifty-five million students in less than three years. According to the National Center for Education Statistics this is more than the 49.8 million students attending public K-12 school in the United States [64]. Students learn a language primarily through completing exercises in solitary lessons and activities. Social features on the platform enable student-student communication, allow students to compete, and give a sense of community. Because learning a foreign language is often a difficult and long-term endeavor, we use various product features to keep students motivated. *The coach* is one feature that has been successful at keeping students engaged. A daily goal is set and the coach keeps track of student progress against a daily goal they set. Students can compare their progress to each other through a social *leaderboard* in which one’s progress is compared to others. Introducing these features in A/B tests and carefully measuring user retention and behavior lets us quantify the impact of product features and changes. Because we have full access to the Duolingo system and instrumentation, we are in a unique position to run experiments and measure outcomes that would otherwise be difficult to perform.

## 1.2 Contributions

A significant contribution of this work is the the very large corpus of data we collected to perform our research. Our dataset serves as the foundation for much of this thesis and has been used in nearly a dozen publications [4, 13, 16, 37, 38, 58, 59, 62, 67, 74, 75] spanning social network analysis, natural language processing, and large-scale algorithms for graph mining. Using this data, we examine several ways in which the social network structure in Twitter influences user behavior:

1. **How do different topics spread in Twitter and what network differences exist across**

**cascades?** Short tokens called *hashtags* are a popular way of attributing a topic to a particular message. For example, discussing political topics such as Health Care Reform or the Grand Old Party can be marked with #HCR and #GOP. These markers allow users to collectively talk about a common topic or event in a concise manner. We study how the 500 most used hashtags spread on Twitter and introduce a model to capture this process. Previously proposed models for information cascades such as the *independent cascade model* and *linear threshold models* don't capture the process in which users become overloaded by a topic, thereby becoming *less* interested in talking about it as it is discussed more. Our probabilistic model is similar to the independent cascade model except that the probability of using the hashtag is a function of how many times a user has already been exposed to the hashtag. These probability curves admit a simple measure for how peaked and sustained the probability is and provides one way of distinguishing between different topics. We also find that across topics there are important network structure conditions that must be met in order for the cascade to grow large.

2. **How does the large 'celebrity' subgraph of Twitter evolve over time?** Twitter's asymmetric follower relationship gives rise to many celebrity figures. At the time this research was conducted, the 1,800 most followed users on Twitter accounted for more than 800M edges in the network. This very large subgraph is an excellent example on which different network evolution models such as preferential attachment could be evaluated. However, Twitter does not provide the times at which social links were created. We devise an algorithm that can accurately estimate the times at which these edges were created. Theoretically this method can be exact and we perform a comprehensive validation experiment to measure its accuracy on celebrity accounts. We find that in practice we can estimate edge creation times that are accurate to within minutes of the actual creation time. Applying the algorithm we are able to evaluate network evolution models, discover the impact of Twitter's signup flow on following behavior, and see how exogenous events such as winning a music award influences network evolution.

We make the following contributions by studying user behavior in Duolingo:

1. **What impact do social features have in an online education platform?** We track several cohorts of students and their use of Duolingo over long periods of time. We find that simply associating the Duolingo account with Facebook does little to change the behavior

of students over time. Frequent ‘nudges’ (receiving notifications) by being followed or by following others significantly increases long-term activity on Duolingo. We further partition these cohorts based on the social features they use to extract correlations between social feature use and long-term learning. These insights will set the direction for future social feature experiments.

2. **What is the network evolution and student-student interaction like in a large, online education platform?** We study the structure of the Duolingo social network (8M+ students, 22M+ edges). Having full control of the system allows us to exactly timestamp and analyze the evolution of the network over time. We also look at how users gain new followers based on their activities on the platform.
3. **What are some indicators that a student will remain committed to learning a foreign language?** Retention is a key metric for Duolingo and many other services. The impact of new features or product changes is evaluated on this metric because it is easily measured and shifts in short-term retention can often be ascertained within one week. This allows for rapid product iteration in a principled manner. We predict long-term student retention based on their behavior shortly after signing up for the service.

# Chapter 2

## Network Structure and the Spread of Hashtags

### 2.1 Background

A growing line of recent research has studied the spread of information on-line, investigating the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have done so [3, 10, 18, 21, 29, 46, 52, 54, 80]. The work in this area has thus far focused primarily on identifying properties that generalize across different domains and different types of information, leading to principles that characterize the process of on-line information diffusion and drawing connections with sociological work on the *diffusion of innovations* [71, 79].

As we begin to understand what is common across different forms of on-line information diffusion, however, it becomes increasingly important to ask about the sources of variation as well. The variations in how different ideas spread is a subject that has attracted the public imagination in recent years, including best-selling books seeking to elucidate the ingredients that make an idea “sticky,” facilitating its spread from one person to another [24, 31]. But despite the fascination with these questions, we do not have a good quantitative picture of how this variation operates at a large scale.

Here are some basic open questions concerning variation in the spread of on-line information.

<sup>0</sup>This chapter is from Romero, Meeder, and Kleinberg [75].

First, the intuitive notion of “stickiness” can be modeled in an idealized form as a probability — the probability that a piece of information will pass from a person who knows or mentions it to another person who is exposed to it. Are simple differences in the value of this probability indeed the main source of variation in how information spreads? Or are there more fundamental differences in the mechanics of how different pieces of information spread? And if such variations exist at the level of the underlying mechanics, can differences in the type or topic of the information help explain them?

**The present work: Variation in the spread of hashtags** In this paper we analyze sources of variation in how the most widely-used hashtags on Twitter spread within its user population. We find that these sources of variation involve not just differences in the probability with which something spreads from one person to another — the quantitative analogue of stickiness — but also differences in a quantity that can be viewed as a kind of “persistence,” the relative extent to which repeated exposures to a piece of information continue to have significant marginal effects on its adoption.

Moreover, these variations are aligned with the topic of the hashtag. For example, we find that hashtags on politically controversial topics are particularly persistent, with repeated exposures continuing to have large relative effects on adoption; this provides, to our knowledge, the first large-scale validation of the “complex contagion” principle from sociology, which posits that repeated exposures to an idea are particularly crucial when the idea is in some way controversial or contentious [14, 15].

Our data is drawn from a large snapshot of Twitter containing large coverage of all tweets during a period of multiple months. From this dataset, we build a network on the users from the structure of interaction via @-messages; for users  $X$  and  $Y$ , if  $X$  includes “@ $Y$ ” in at least  $t$  tweets, for some threshold  $t$ , we include a directed edge from  $X$  to  $Y$ . @-messages are used on Twitter for a combination of communication and name-invocation (such as mentioning a celebrity via @, even when there is no expectation that they will read the message); under all these modalities, they provide evidence that  $X$  is paying attention to  $Y$ , and with a strength that can be tuned via the parameter  $t$ .<sup>1</sup>

<sup>1</sup>One can also construct a directed network from the *follower* relationship, including an edge from  $X$  to  $Y$  if  $X$  follows  $Y$ . We focus here on @-messages in part because of a data resolution issues — they can be recovered with exact time stamps from the tweets themselves — but also because of earlier research suggesting that users often follow other users in huge numbers and hence potentially less discriminately, whereas interaction via @-messages



For a given user  $X$ , we call the set of other users to whom  $X$  has an edge the *neighbor set* of  $X$ . As users in  $X$ 's neighbor set each mention a given hashtag  $H$  in a tweet for the first time, we look at the probability that  $X$  will first mention it as well; in effect, we are asking, "How do successive exposures to  $H$  affect the probability that  $X$  will begin mentioning it?" Concretely, following the methodology of [18], we look at all users  $X$  who have not yet mentioned  $H$ , but for whom  $k$  neighbors have; we define  $p(k)$  to be the fraction of such users who mention  $H$  before a  $(k + 1)^{\text{st}}$  neighbor does so. In other words,  $p(k)$  is the fraction of users who adopt the hashtag directly after their  $k^{\text{th}}$  "exposure" to it, given that they had not yet adopted it.

As an example, Figure 2.1 shows a plot of  $p(k)$  as a function of  $k$  averaged over the 500 most-mentioned hashtags in our dataset. Note that these top hashtags are used in sufficient volume that one can also construct meaningful  $p(k)$  curves for each of them separately, a fact that will be important for our subsequent analysis. For now, however, we can already observe two basic features of the average  $p(k)$  curve's shape: a ramp-up to a peak value that is reached relatively early (at  $k = 2, 3, 4$ ), followed by a decline for larger values of  $k$ . In keeping with the informal discussion above, we define the *stickiness* of the curve to be the maximum value of  $p(k)$  (since this is the maximum probability with which an exposure to  $H$  transfers to another user), and the *persistence* of the curve to be a measure of its rate of decay after the peak.<sup>2</sup> We will find that, in a precise sense, these two quantities — stickiness and persistence — are sufficient to approximately characterize the shapes of individual  $p(k)$  curves.

**Variation in Adoption Dynamics Across Topics** The shape of  $p(k)$  averaged over all hashtags is similar to analogous curves measured recently in other domains [18], and our interest here is in going beyond this aggregate shape and understanding how these curves vary across different kinds of hashtags. To do this, we first classified the 500 most-mentioned hashtags according to their topic. We then average the curves  $p(k)$  separately within each category and compare their shapes.<sup>3</sup>

indicates a kind of attention that is allocated more parsimoniously, and with a strength that can be measured by the number of repeat occurrences [33].

<sup>2</sup>We formally define persistence in Section 2.3; roughly, it is the ratio of the area under the curve to the area of the largest rectangle that can be circumscribed around it.

<sup>3</sup>In Section 2.2 we describe the methodology used to perform this manual classification in detail. In brief, we compared independent classifications of the hashtags obtained by disjoint means, involving annotation by the authors compared with independent annotation by a group of volunteers. Our results based on the average curves arising from this classification are robust in the following sense: despite differences in classification of some individual hashtags by the two groups, the curves themselves exhibit essentially identical behavior when computed from either

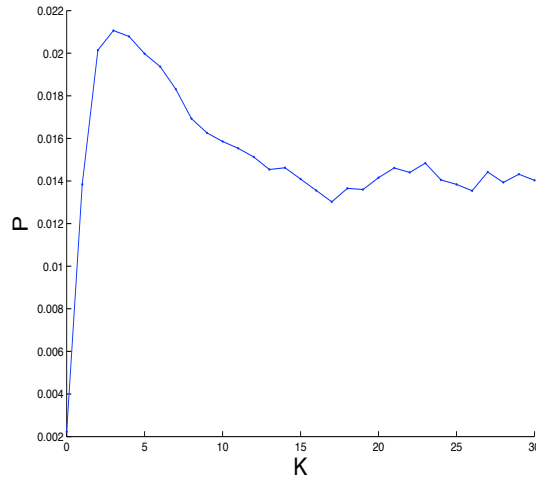


Figure 2.1: Average exposure curve for the top 500 hashtags.  $P(K)$  is the fraction of users who adopt the hashtag directly after their  $k^{\text{th}}$  exposure to it, given that they had not yet adopted it

Many of the categories have  $p(k)$  curves that do not differ significantly in shape from the average, but we find unusual shapes for several important categories. First, for political hashtags, the persistence has a significantly larger value than the average — in other words, successive exposures to a political hashtag have an unusually large effect relative to the peak. This is striking in the way that it accords with the “complex contagion” principle discussed earlier: when a particular behavior is controversial or contentious, people may need more exposure to it from others before adopting it themselves [14, 15].

In contrast, we find a different form of unusual behavior from a class of hashtags that we refer to as Twitter *idioms* — a kind of hashtag that will be familiar to Twitter users in which common English words are concatenated together to serve as a marker for a conversational theme (e.g. #cantlivewithout, #dontyouhate, #iloveitwhen, and many others, including concatenated markers for weekly Twitter events such as #musicmonday and #followfriday.) Here the stickiness is high, but the persistence is unusually low; if a user doesn’t adopt an idiom after a small number of exposures, the marginal chance they do so later falls off quickly.

**Subgraph Structure and Tie Strength** In addition to the person-to-person mechanics of spread, it is also interesting to look at the overall structure of interconnections among the initial adopters of a hashtag. To do this, we take the first  $m$  individuals to mention a particular hashtag  $H$ , and of the two classifications separately, as well as from an intersection of the two classifications.

we study the structure of the subgraph  $G_m$  induced on these first  $m$  mentioners. In this structural context, we again find that political hashtags exhibit distinctive features — in particular, the subgraphs  $G_m$  for political hashtags  $H$  tend to exhibit higher internal degree, a greater density of triangles, and a large number of nodes not in  $G_m$  who have significant numbers of neighbors in it. This is again broadly consistent with the sociological premises of complex contagion, which argues that the successful spread of controversial behaviors requires a network structure with significant connectivity and significant local clustering.

Within these subgraphs, we can consider a set of sociological principles that are related to complex contagion but distinct from it, centered on the issue of *tie strength*. Work of McAdam and others has argued that the sets of early adopters of controversial or risky behaviors tend to be rich in strong ties, and that strong ties are crucial for these activities [56, 57] — in contrast to the ways in which learning about novel information can correspondingly benefit from transmission across weaker ties [27].

When we look at tie strength in these subgraphs, we find a somewhat complex picture. Because subgraphs  $G_m$  for political hashtags have significantly more edges, they have more ties of all strengths, including strong ties (according to several different definitions of strength summarized in Section 2.4). This aspect of the data aligns with the theories of McAdam and others. However, the fraction of strong ties in political subgraphs  $G_m$  is actually *lower* than the fraction of strong ties for the full population of widely-used hashtags, indicating the overall greater density of edges in political subgraphs comes more dominantly from a growth in weak ties than from strong ones. The picture that emerges of early-adopter subgraphs for political hashtags is thus a subtle one: they are structures whose communication patterns are more densely connected than the early-adopter subgraphs for other hashtags, and this connectivity comes from a core of strong ties embedded in an even larger profusion of weak ties.

**Interpreting the Findings** When we look at politically controversial topics on Twitter, we therefore see both direct reflections and unexpected variations on the sociological theories concerning how such topics spread. This is part of a broader and important issue: understanding differences in the dynamics of contentious behavior in the off-line world versus the on-line world. It goes without saying that the use of a hashtag on Twitter isn't in any sense comparable, in terms of commitment or personal risk, to taking part in activism in the physical world (a point

recently stressed in a much-circulated article by Malcolm Gladwell [25]). But the underlying issue persists on Twitter: political hashtags are still riskier to use than conversational idioms, albeit at these much lower stakes, since they involve publicly aligning yourself with a position that might alienate you from others in your social circle. The fact that we see fundamental aspects of the same sociological principles at work both on-line and off-line suggests a certain robustness to these principles, and the differences that we see suggest a perspective for developing deeper insights into the relationship between these behaviors in the on-line and off-line domains.

This distinction between contentious topics in the on-line and off-line worlds is one issue to keep in mind when interpreting these results. Another is the cumulative nature of the findings. As with any analysis at this scale, we are not focusing on why any one individual made the decisions they did, nor is it the case that that Twitter users are even aware of all the tweets containing their exposures to hashtags via neighbors. Rather, the point is that we still find a strong signal in an aggregate sense — as a whole, the population is exhibiting differences in how it responds to hashtags of different types, and in ways that accord with theoretical work in other domains.

A further point to emphasize is that our focus in this work is on the hashtags that succeeded in reaching large numbers of people. It is an interesting question to consider what distinguishes a hashtag that spreads widely from one that fails to attract attention, but that is not the central question we consider here. Rather, what we are identifying is that among hashtags that do reach many people, there can nevertheless be quite different mechanisms of contagion at work, based on variations in stickiness and persistence, and that these variations align in interesting ways with the topic of the hashtag itself.

**Simulated Spreading** Finally, an interesting issue here is the interaction between the  $p(k)$  curve and the subgraph  $G_m$  for a given hashtag  $H$  — clearly the two develop in a form of co-evolution, since the addition of members via the curve  $p(k)$  determines how the subgraph of adopters takes shape, but the structure of this subgraph — particularly in the connections between adopters and non-adopters — affects who is likely to use the hashtag next. To understand how  $p(k)$  and  $G_m$  relate to each other, it is natural to consider questions of the following form: how would the evolution of  $G_m$  have turned out differently if a different  $p(k)$  curve had been in effect? Or correspondingly, how effectively would a hashtag with curve  $p(k)$  have spread if it had started from a different subgraph  $G_m$ ? Clearly it is difficult to directly perform this counterfactual

experiment as stated, but we obtain insight into the structure of the question by simulating the  $p(k)$  curve of each top hashtag on the subgraph  $G_m$  of each other top hashtag. In this way, we begin to identify some of the structural factors at work in the interplay between the mechanics of person-to-person influence and the network on which it is spreading.

## 2.2 Dataset, Network Definition, and Hashtag Classification

**Data Collection and Network Definition** From August 2009 until January 2010 we crawled Twitter using their publicly available API. Twitter provides access to only a limited history of tweets through the search mechanism; however, because user identifiers have assigned contiguously since an early point in time, we simply crawled each user in this range. Due to limitations of the API, if a user has more than 3,200 tweets we can only recover the last 3,200 tweets; all messages of any user with fewer than this many tweets are available. We collected over three billion messages from more than 60 million users during this crawl.

As discussed in Section 2.1, in addition to extracting tweets and hashtags within them, we also build a network on the users, connecting user  $X$  to user  $Y$  if  $X$  directed at least  $t$  @-messages to  $Y$ . In our analyses we use  $t = 3$ , except when we are explicitly varying this parameter. The resulting network contains 8,509,140 non-isolated nodes and 50,814,366 links. As noted earlier, there are multiple ways of defining a network on which hashtags can be viewed as diffusing, and our definition is one way of defining a proxy for the attention that users  $X$  pay to other users  $Y$ .

**Hashtag Selection and Classification** To create a classification of hashtags by category, we began with the 500 hashtags in the data that had been mentioned by the most users. From manual inspection of this list, we identified eight broad categories of hashtags that each had at least 20 clear exemplars among these top hashtags, and in most cases significantly more. (Of course, many of the top 500 hashtags fit into none of the categories.) We formulated definitions of these categories as shown in Table 2.1. Then we applied multiple independent mechanisms for classifying the hashtags according to these categories. First, the authors independently annotated each hashtag, and then had a reconciliation phase in which they noted errors and arrived at a majority judgment on each annotation. Second, the authors solicited a group of independent annotators, and took the majority among their judgments. Annotators were provided with the

Category	Definition
Celebrity	The name of a person or group (e.g. music group) that is featured prominently in entertainment news. Political figures or commentators with a primarily political focus are not included. The name of the celebrity may be embedded in a longer hashtag referring to some event or fan group that involves the celebrity. Note that many music groups have unusual names; these still count under the “celebrity” category.
Games	Names of computer, video, MMORPG, or twitter-based games, as well as groups devoted to such games.
Idiom	A tag representing a conversational theme on twitter, consisting of a concatenation of at least two common words. The concatenation can’t include names of people or places, and the full phrase can’t be a proper noun in itself (e.g. a title of a song/movie/organization). Names of days are allowed in the concatenation, because of the the Twitter convention of forming hashtags involving names of days (e.g. MusicMonday). Abbreviations are allowed only if the full form also appears as a top hashtag (so this rules out hashtags including omg, wtf, lol, nsfw).
Movies/TV	Names of movies or TV shows, movie or TV studios, events involving a particular movie or TV show, or names of performers who have a movie or TV show specifically based around them. Names of people who have simply appeared on TV or in a movie do not count.
Music	Names of songs, albums, groups, movies or TV shows based around music, technology designed for playing music, or events involving any of these. Note that many music groups have unusual names; these still count under the “music” category.
Political	A hashtag that in your opinion often refers to a politically controversial topic. This can include a political figure, a political commentator, a political party or movement, a group on twitter devoted to discussing a political cause, a location in the world that is the subject of controversial political discussion, or a topic or issue that is the subject of controversial political discussion. Note that this can include political hashtags oriented around countries other than the U.S.
Sports	Names of sports teams, leagues, athletes, particular sports or sporting events, fan groups devoted to sports, or references to news items specifically involving sports.
Technology	Names of Web sites, applications, devices, or events specifically involving any of these.

Table 2.1: Definitions of categories used for annotation.

Category	Examples	Category	Examples
Celebrity	mj, brazilwantsjb, regis, iwantpeterfacinelli	Music	thisiswar, mj, musicmonday, pandora
Games	mafiawars, spymaster, mw2, zyangapirates	Political	tcot, glennbeck, obama, hcr
Idiom	cantlivewithout, dontyouhate, musicmonday	Sports	golf, yankees, nhl, cricket
Movies/TV	lost, glennbeck, bones, newmoon	Technology	digg, iphone, jquery, photoshop

Table 2.2: A small set of examples of members in each category.

category definitions, and for each hashtag were provided with the tag’s definitions (when present) from the Web resources Wthashtag and Tagalus, as well as links to Google and Twitter search results on the tag. Finally, since the definition of the “idiom” category is purely syntactic, we did not use annotators for this task, but only for the other seven categories.

Clearly even with this level of specificity, involving both human annotation and Web-based definitional resources, there are ultimately subjective judgments involved in category assignments. However, given the goal of understanding variations in hashtag behavior across topical categories, at some point in the process a set of judgments of this form is unavoidable. What we find is the results are robust in the presence of these judgments: the level of agreement among annotators was uniformly high, and the plots presented in the subsequent sections show essentially identical behavior regardless of whether they are based on the authors’ annotations, the independent volunteers’ annotations, or the intersection of the two. To provide the reader with some intuition for the kinds of hashtags that fit each category, we present a handful of illustrative examples in Table 2.2, drawn from the much larger full membership in each category. The full

category memberships can be seen at <http://www.cam.cornell.edu/~dromero/top500ht>.

## 2.3 Exposure Curves

**Basic definitions** In order to investigate the mechanisms by which hashtag usage spreads among Twitter users, we begin by reviewing two ways of measuring the impact that exposure to others has in an individual's choice to adopt a new behavior (in this case, using a hashtag) [18]. We say that a user is *k*-exposed to hashtag *h* if he has not used *h*, but has edges to *k* other users who have used *h* in the past. Given a user *u* that is *k*-exposed to *h* we would like to estimate the probability that *u* will use *h* in the future. Here are two basic ways of doing this.

**Ordinal time estimate.** Assume that user *u* is *k*-exposed to some hashtag *h*. We will estimate the probability that *u* will use *h* before becoming (*k* + 1)-exposed. Let  $E(k)$  be the number of users who were *k*-exposed to *h* at some time, and let  $I(k)$  be the number of users that were *k*-exposed and used *h* before becoming (*k* + 1)-exposed. We then conclude that the probability of using the hashtag *h* while being *k*-exposed to *h* is  $p(k) = \frac{I(k)}{E(k)}$ .

**Snapshot estimate.** Given a time interval  $T = (t_1, t_2)$ , assume that a user *u* is *k*-exposed to some hashtag *h* at time  $t = t_1$ . We will estimate the probability that *u* will use *h* sometime during time interval *T*. We let  $E(k)$  be the number of users who were *k*-exposed to *h* at time  $t = t_1$ , and let  $I(k)$  be the number of users who were *k*-exposed to *h* at time  $t = t_1$  and used *h* sometime before  $t = t_2$ . We then conclude that  $p(k) = \frac{I(k)}{E(k)}$  is the probability of using *h* before time  $t = t_2$ , conditioned on being *k*-exposed to *h* at time  $t = t_1$ . We will refer to  $p(k)$  as an *exposure curve*; we will also informally refer to it as an *influence curve*, although it is being used only for prediction, not necessarily to infer causal influence.

The ordinal time approach requires more detailed data than the snapshot method. Since our data are detailed enough that we are able to generate the ordinal time estimate, we only present the results based on the ordinal time approach; however, we have confirmed that the conclusions hold regardless of which approach is followed. This is not surprising since it has been argued that sufficiently many snapshot estimates contain enough information to infer the the ordinal time estimate [18].

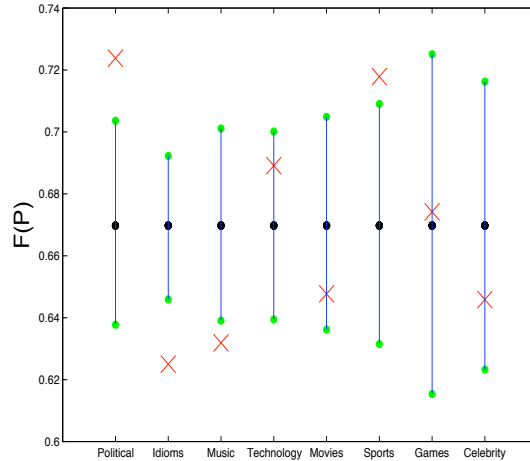


Figure 2.2:  $F(P)$  for the different types of hashtags. The black dots are the average  $F(P)$  among all hashtags, the red x is the average for the specific category, and the green dots indicate the 90% expected interval where the average for the specific set of hashtags would be if the set was chosen at random. Each point is the average of a set of at least 10 hashtags

**Comparison of Hashtag Categories: Persistence and Stickiness** We calculated ordinal time estimates  $P(k)$  for each one of the 500 hashtags we consider. For each point on each curve we calculate the 95% Binomial proportion confidence interval. We observed some qualitative differences between the curves corresponding to different hashtags. In particular, we noticed that some curves increased dramatically initially as  $k$  increased but then started to decrease relatively fast, while other curves increased at a much slower rate initially but then saturated or decreased at a much slower rate. As an example, Figure 2.3 shows the influence curves for the hashtags #cantlivewithout and #hcr. We also noticed that some curves had much higher maximum values than others.<sup>4</sup>

In this discussion, we are basing differences among hashtags on different structural properties of their influence curves. In order to make these distinctions more precise we use the following measures.

First, we formalize a notion of “persistence” for an influence curve, capturing how rapidly it decays. Formally, given a function  $P : [0, K] \rightarrow [0, 1]$  we let  $R(P) = K \max_{k \in [0, K]} \{P(k)\}$  be the area of the rectangle with length  $K$  and height  $\max_{k \in [0, K]} \{P(k)\}$ . We let  $A(P)$  be the area under the

<sup>4</sup>As  $k$  gets larger the amount of data used to calculate  $P(k)$  decreases, making the error intervals very large and the curve very noisy. In order to take this into account we only defined  $P(k)$  when the relative error was less than some value  $\theta$ . Throughout the study we checked that the results held for different values of  $\theta$ .



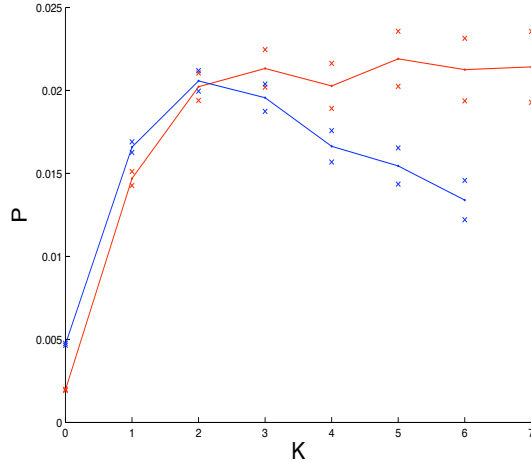


Figure 2.3: Sample exposure curves for hashtags #cantlivewithout (blue) and #hcr (red).

curve  $P$  assuming the point  $P(k)$  is connected to the point  $P(k + 1)$  by a straight line. Finally, we let  $F(P) = \frac{A(P)}{R(P)}$  be the *persistence* parameter.

When an influence curve  $P$  initially increases rapidly and then decreases, it will have a smaller value of  $F(P)$  than a curve  $\tilde{P}$  which increases slowly and then saturates. Similarly, an influence curve  $P$  that slowly increases monotonically will have a smaller value of  $F(P)$  than a curve  $\tilde{P}$  that initially increases rapidly and then saturates. Hence the measure  $F$  captures some differences in the shapes of the influence curves. In particular, applying this measure to an influence curve would tell us something about its persistence; the higher the value of  $F(P)$ , the more persistent  $P$  is.

Second, given an influence curve  $P : [0, K] \rightarrow [0, 1]$  we let  $M(P) = \max_{k \in [0, K]} \{P(k)\}$  be the *stickiness* parameter, which gives us a sense for how large the probability of usage can be for a particular hashtag based on the most effective exposure.

We are interested in finding differences between the spreading mechanism of different topics on Twitter. We start by finding out if hashtags corresponding to different topics have influence curves with different shapes. We found significant differences in the values of  $F(P)$  for different topics. Figure 2.2 shows the average  $F(P)$  for the different categories, compared to a baseline in which we draw a set of categories of the same size uniformly at random from the full collection of 500. We see that politics and sports have an average value of  $F(P)$  which is significantly higher than expected by chance, while for Idioms and Music it is lower. This suggests that the mechanism that controls the spread of hashtags related to sports or politics tends to be more per-

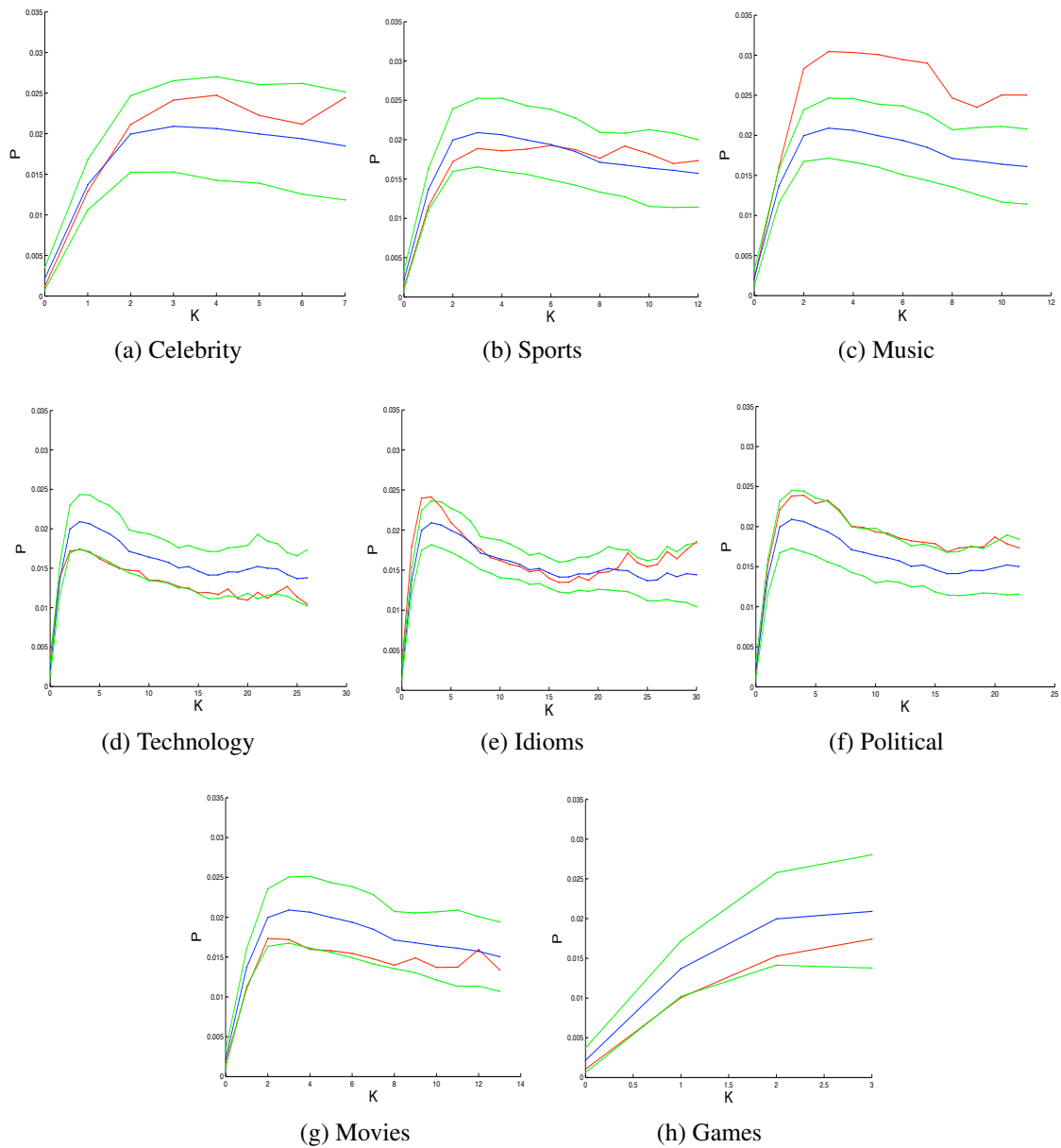


Figure 2.4: Point-wise average influence curves. The blue line is the average of all the influence curves, the red line is the average for the set of hashtags of the particular topic, and the green lines indicate the interval where the red line is expected to be if the hashtags were chosen at random.

sistent than average; repeated exposures to users who use these hashtags affects the probability that a person will eventually use the hashtag more positively than average. On the other hand, for Idioms and Music, the effect of repeated exposures falls off more quickly, relative to the peak, compared to average.

Figure 2.4 shows the point-wise average of the influence curves for each one of the categories. Here we can see some of the differences in persistence and stickiness the curves have. For example, the stickiness of the topics Music, Celebrity, Idioms, and politics tends to be higher than average since the average influence curve for those categories tends to be higher than the average influence curve for all hashtags, while that of Technology, Movies, and Sports tends to be lower than average. On the other hand, these plots give us more intuition on why we found that politics and Sports have a high persistence while for Idioms and Music it is low. In the case of Politics, we see that the red curve starts off just below the green curve (the upper error bar) and as  $k$  increases, the red curve increases enough to be above the green. Similarly, the red curve for Sports starts below the blue curve and it ends above it. In the case of Idioms, the red curve initially increases rapidly but then it drops below the blue curve. Similarly, the red curve for Music is always very high and above all the other curves, but it drops faster than the other curves at the end.

**Approximating Curves via Stickiness and Persistence** When we compare curves based on their stickiness and persistence, it is important to ask whether these are indeed an adequate pair of parameters for discussing the curves’ overall “shapes.” We now establish that they are, in the following sense: we show that these two parameters capture enough information about the influence curves that we can approximate the curves reasonably well given just these two parameters. Assume that for some curve  $P$  we are given  $F(P)$  and  $M(P)$ . We will also assume that we know the maximum value of  $k = K$  for which  $P(k)$  is defined. Then we will construct an approximation curve  $\tilde{P}$  in the following way:

1. Let  $\tilde{P}(0) = 0$
2. Let  $\tilde{P}(2) = M(P)$
3. Now we will let  $\tilde{P}(K)$  be such that  $F(\tilde{P}) = F(P)$ . This value turns out to be  $\tilde{P}(K) = \frac{M(P) * K * (2 * F(P) - 1)}{K - 2}$
4. Finally, we will make  $\tilde{P}$  be piecewise linear with one line connecting the points  $(0, 0)$  and

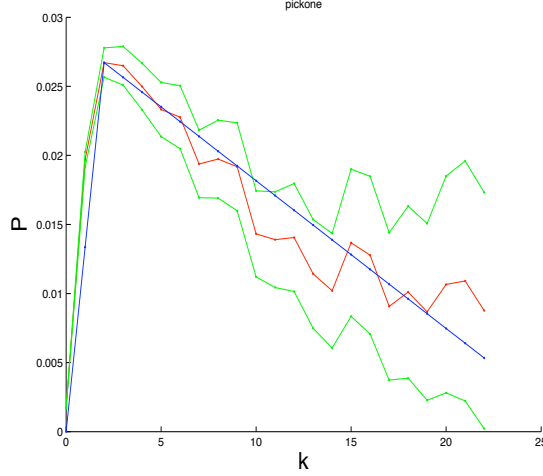


Figure 2.5: Example of the approximation of an influence curve. The red curve is the influence curve for the hashtag #pickone, the green curves indicate the 95% binomial confidence interval, and the blue curve is the approximation.

$(2, M(P))$ , and another line connecting the points  $(2, M(P))$  and  $(K, \frac{M(P)*K*(2*F(P)-1)}{K-2})$ .

Figure 2.5 shows an example of an approximation for a particular influence curve. In order to test the quality of the approximation  $\tilde{P}$  we define the approximation error between  $\tilde{P}$  and  $P$  as the mean absolute error

$$E(P, \tilde{P}) = \frac{1}{K} \sum_{k=0}^K |P(k) - \tilde{P}(k)|$$

and compare it with the mean absolute of the error  $E(P)$  obtained from the 95% confidence intervals around each point  $P(k)$ . The average approximation error among all the influence curves is 0.0056 and the average error of based on the confidence intervals is 0.0050. The approximation error is slightly smaller, which means that our approximation is, on average, within the 95% confidence interval from the actual influence curve. This suggests the information contained in the stickiness and persistence parameters are enough to accurately approximate the influence curves and gives more meaning to the approach of comparing the curves by comparing these two parameters.

**Frequency of Hashtag Usage** We have observed that different topics have differences in their spreading mechanisms. We also found that they differed in other ways. For example, we see

Type	Mdn. Mentions	Mdn. Users	Mdn. Ment./User
All HTS	93,056	15,418	6.59
Political	132,180	13,739	10.17
Sports	98,234	11,329	9.97
Idioms	99,317	26,319	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

Table 2.3: Median values for number of mentions, number of users, and number of mentions per user for different types of hashtags

some variation in the number of mentions and the number of users of each category. Table 2.3 shows the different median values for number of mentions, number of users, and number of mentions per user for different types of hashtags. We see that while Idioms and Technology hashtags are used by many users compared to others, each user only uses the hashtag a few times and hence the total number of mentions of these categories is not much higher than others. On the other hand, only relatively few people used Political and Games hashtags, but each one of them used them many times, making them the most mentioned categories. In the case of games, a contributing factor is that some of users of game hashtags allow external websites to post on their Twitter account every time they accomplish something in the game, which tends to happen very often. It is not clear that there is a correspondingly simple explanation for the large number of mentions per user for political hashtags, but one can certainly conjecture that it may reflect something about the intensity with which these topics are discussed by the users who engage in such discussions; this is an interesting issue to explore further.

## 2.4 The structure of initial sets

The spread of a given piece of information is affected by the diffusion mechanism controlled by the influence curves discussed in the previous section, but it may also be affected by the structure of the network relative to the users of the hashtag. To explore this further, we looked at the subgraph  $G_m$  induced by the first  $m$  people who used a given hashtag. We found that there are important differences in the structure of those graphs.

In particular, we consider differences in the structures of the subgraphs  $G_m$  across different categories. For each graph  $G_m$ , across all hashtags and a sequence of values of  $m$ , we compute several structural parameters. First, we compute the average degree of the nodes and the number of triangles in the graph. Then, we defined the *border* of  $G_m$  to be the set of all nodes not in  $G_m$  who have at least one edge to a node in  $G_m$ , and we define the *entering degree* of a node in the border to be the number of neighbors it has in  $G_m$ . We consider the size of the border and the average entering degree of nodes in the border.

Looking across all categories, we find that political hashtags are the category in which the most significant structural differences from the average occur. Table 2.4 shows the averages for political hashtags compared to the average for all hashtags, using the subgraphs  $G_{500}$  on the first 500 users.<sup>5</sup> In brief, the early adopters of a political hashtag message with more people, creating more triangles, and with a border of people who have more links on average into the early adopter set. The number of triangles, in fact, is high even given the high average degree; clearly one should expect a larger number of triangles in a subgraph of larger average degree, but in fact the triangle count for political hashtags is high even when compared against a baseline consisting of non-political hashtags with comparable average degrees. These large numbers of edges and triangles are consistent with the predictions of complex contagion, which argues that such structural properties are important for the spread of controversial topics [15].

**Tie Strength** There is an interesting further aspect to these structural results, obtained by looking at the *strength* of the ties within these subgraphs. There are multiple ways of defining tie strength from social media data [23], and here we consider two distinct approaches. One approach is to use the total number of @-messages sent across the link as a numerical measure of strength. Alternately, we can declare a link to be strong if and only if it is *reciprocated* (i.e. declaring  $(X, Y)$  to be strong if and only if  $(Y, X)$  is in the subgraph as well, following a standard working notion of reciprocation as a proxy for tie strength in the sociology literature [28]).

Under both definitions, we find that the fraction of strong ties in subgraphs  $G_m$  for political hashtags is in fact significantly lower than the fraction of strong ties in subgraphs  $G_m$  for our set of hashtags overall. However, since political subgraphs  $G_m$  contain so many links relative to the typical  $G_m$ , we find that they have a larger absolute number of strong ties. As noted in

<sup>5</sup>The results are similar for  $G_m$  with a range of other values of  $m \neq 500$ .

Type	I	II	III	IV
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

Table 2.4: Comparison of graphs induced by the first 500 early adopters of political hashtags and average hashtags. Column definitions: I. Average degree, II. Average triangle count, III. Average entering degree of the nodes in the border of the graphs, IV. Average number of nodes in the border of the graphs. The error bars indicate the 95% confidence interval of the average value of a randomly selected set of hashtags of the same size as Political.

the introduction, standard sociological theories suggest that we should see many strong ties in subgraphs  $G_m$  for political topics, but the picture we obtain is more subtle in that the growth in strong ties comes with an even more significant growth in weak ties. Understanding these competing forces in the structural behavior of such subgraphs is an interesting open question.

## 2.5 Simulations

We have observed that for some hashtags, such as those relating to political subjects, users are particularly affected by multiple exposures before using them. We also know that the subgraphs on which political hashtags initially spread have high degrees and extensive clustering. To what extent do these aspects intrinsically go together? Do these types of political hashtags spread effectively because of the close-knit network of the initial users? Are political subjects less likely to successfully spread on sparsely connected initial sets?

In this section, we try to obtain some initial insight into these questions through a simulation model — not only in the context of political hashtags but also in the context of the other categories. In particular, we develop a model that naturally complements the process used to calculate the  $p(k)$  functions. We perform simulations of this model using the measured  $p(k)$  functions and a varying number of the first users who used each hashtag on the actual influence network. Additionally, we record the progression of the cascade and track its spread through the network. By trying the  $p(k)$  curve of a hashtag on the initial sets of other hashtags, and by varying the size of the initial sets, we can gain insight into the factors that lead to wide-spreading cascades.

### 2.5.1 The Simulated Model

We wish to simulate cascades using the measured  $p(k)$  curves, the underlying network of users, and in particular the observed subgraphs  $G_m$  of initial adopters. In this discussion, and in the model we present hereafter, we refer to the moment at which a node adopts a hashtag as its *activation*. We operationalize the model implicit in the definition of the function  $p(k)$ , leading to the following natural simulation process on a graph  $G = (V, E)$ .

First, we activate all nodes in the starting set  $I$ , and mark them all as newly active. In a general iteration  $t$  (starting with  $t = 0$ ), we will have a currently active set  $A_t$  and a subset  $N_t \subseteq A_t$  of *newly active* nodes. (In the opening iteration, we have  $A_0 = N_0 = I$ .) Newly active nodes have an opportunity to activate nodes  $u \in V - A_t$ , with the probabilities of success on  $u$  determined by the  $p(k)$  curve and the number of nodes in  $A_t - N_t$  who have already tried and failed to activate  $u$ .

Thus, we consider each node  $u \in V - A_t$  that is a neighbor of at least one node in  $N_t$ , and hence will experience at least one activation attempt. Let  $k_t(u)$  be the number of nodes in  $A_t - N_t$  adjacent to  $u$ ; these are the nodes that have already tried and failed to activate  $u$ . Let  $\Delta_t(u)$  be the number of nodes in  $N_t$  adjacent to  $u$ . Each of these neighbors in  $N_t$  will attempt to activate  $u$  in sequence, and they will succeed with probabilities  $p(k_t(u) + 1), p(k_t(u) + 2), \dots, p(k_t(u) + \Delta_t(u))$ , since these are the success probabilities given the number of nodes that have already tried and failed to activate  $u$ . At the end, we define  $N_{t+1}$  to be the set of nodes  $u$  that are newly activated by the attempts in this iteration, and  $A_{t+1} = A_t \cup N_{t+1}$ .

### 2.5.2 Simulation Results

We simulate how a cascade that spreads according to the  $p(k)$  curve for some hashtag evolves when seeded with an initially active user set of some other hashtag. In total, there are 250,000 ( $p(k)$ , start set) hashtag combinations we examine. We additionally vary the size of the initially active set to be 100, 500, or 1,000 users. Since we want to study how a hashtag blossoms from being used by a few starting nodes to a large number of users, we must be careful about how we select the size of our starting sets. We believe that these initial set sizes capture the varying topology observed in Section 2.4 and are not too large as to guarantee wide-spreading cascade. For 100 and 500 starting nodes we run five simulations on each ( $p(k)$ , start set) pair, and for



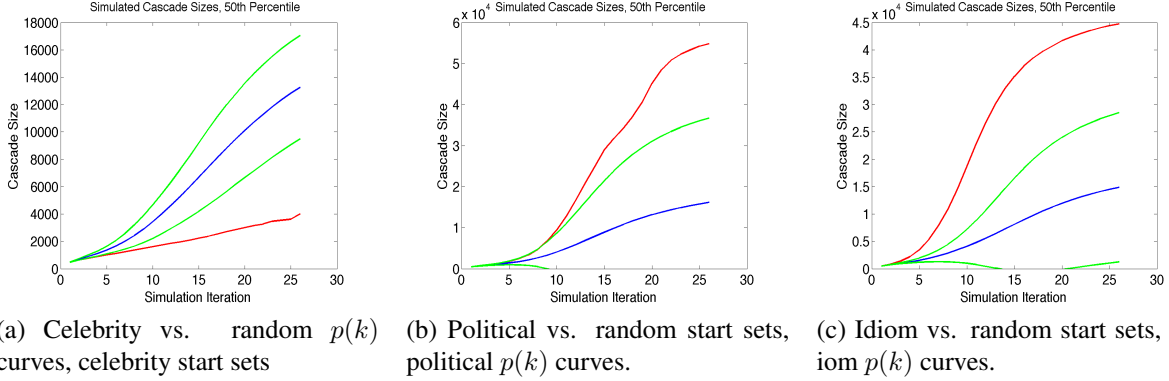


Figure 2.6: **Validating Category Differences:** The median cascade sizes for three different categories. In 2.6a we show the difference between celebrity and random  $p(k)$  curves on celebrity starting sets. celebrity hashtags and random . In 2.6b we show politics. Finally, we show idioms in 2.6c. All starting sets consist of 500 users.

1,000 starting nodes we run only two simulations.

The simulation is instrumented at each iteration; we record the size of the cascade, the number of nodes influenced by active users, and the number of inactive users influenced by active users. Furthermore, each simulation runs for at most 25 iterations. We found that this number of iterations was large enough to observe interesting variation in cascade sizes yet still be efficiently simulated.

We calculate the mean and the 5th, 10th, ..., 90th, 95th percentiles of cascade sizes after each iteration. For each category, we measure these twenty measures based on all of the simulations where the  $p(k)$  hashtag and the starting set hashtag are both chosen from the category. We then compare these measurements to the results when a random set of hashtags is used to decide the  $p(k)$  curve, the starting set, or both the  $p(k)$  curve and the starting set. The cardinality of this random set is the same as the number of hashtags in the category. We sample these random choices 10,000 times to estimate the distribution of these measured features.

Using these samples, we test the measurements for statistical significance. In particular, we look at how the ‘category’ cascades (those in which both hashtag choices are from the category set) compare to cascades in which the  $p(k)$  curve or starting set hashtags were chosen randomly. In all of the following figures, the red line indicates the value of the measurements over the set of simulations in which  $p(k)$  curve and the start set come from category hashtags. The blue line is the average feature measurement over the random choices, and the green lines specify

two standard deviations from the mean value. The cascade behavior of a category is statistically significant with respect to one of the measured features when most of the red curve lies outside of the region between the two green curves.

We compare how the  $p(k)$  curves for a category perform on start sets from the same category and on random start sets. We additionally evaluate how random  $p(k)$  curves and category  $p(k)$  curves perform on category start sets. In general, categories either performed below or above the random sets in both of these measures. Some particular observations are

- **Celebrities and Games:** Compared to random starting sets, we find that start sets from these categories generate smaller cascades when the  $p(k)$  curves are chosen from their respective categories. This difference is statistically significant.
- **Political and Idioms:** These categories'  $p(k)$  curves and start sets perform better than a random choice. This is especially true for the smaller cascades (5 - 30th percentiles).
- **Music:** This category is interesting because the music  $p(k)$  curves perform better than random  $p(k)$  curves on music starting sets, music  $p(k)$  curves perform better on random starting sets than on music starting sets, regardless of the number of initially active users. This is the only category in which the  $p(k)$  and start set 'goodness' differs.
- **Movies, Sports, and Technology:** These categories don't exhibit particularly strong over or underperformance compared a random choice of  $p(k)$  hashtags and starting set hashtags.

# Chapter 3

## The Evolution of Twitter's Social Graph

### 3.1 Background

Twitter is a popular social networking website that enables users to send and receive short messages of at most 140 characters, which are also called *tweets*. Tweets are not highly directed messages like email, but are instead broadcast to all of a user's *followers*. Following is the sole social connection in Twitter; a user's primary view in Twitter is a reverse chronological stream of tweets from accounts that user is following. Academic studies of Twitter typically represent the user population as a directed graph (or directed network) because the following relationship can, and often is, asymmetric.

Information about Twitter can be gathered from the open Twitter API [1] which provides access to a broad range of information including both tweet content and the current social graph. Despite providing timing information on most other accessible data, Twitter does not provide the time at which edges in the social network were formed. Unlike smaller social networks for which recrawling or continuous observation of the social graph is feasible, even a single crawl of a relatively small fraction of Twitter can be a time-consuming enterprise. A major contribution of this work is a simple method to assign times to the creation of edges that only requires one static social network snapshot. For any edge in the network, the assigned time is a *lower bound* for the time the edge was created. Despite only using one crawl, for users who rapidly gain followers, we show that the process of assigning times, which we call *timestamping*, can be extremely accurate both in theory and in practice.

<sup>0</sup>This chapter is from Meeder et al [58]

Fortunately, Twitter has interesting users who rapidly gain followers. Most users only have a few followers but some accounts on Twitter have garnered an enormous number of followers. These popular accounts include real-life celebrities such as Lady Gaga and Justin Bieber, politicians such as President Barack Obama and former vice president Al Gore, and news media such as CNN Breaking News and The New York Times. These users can gain thousands of new followers a day on Twitter. Twitter itself promotes following others by presenting a list of recommended users, also called the *suggested users list*, at the last step of creating an account. A new user is encouraged to follow these suggested users as an introduction to Twitter and the lucky users placed on the suggested users list gain elevated numbers of followers per day.

How does the rate of accumulation of followers change over time for these prominent users in Twitter? What are the key factors that influence these changes? What is the pattern of users following celebrities in relation to their account creation times and can this pattern for existing users tell us anything about the importance of celebrities to the Twitter graph? For all these questions, we need accurate temporal information about edge formation times that is not available from Twitter's API. We present a simple timestamping method that recovers estimates of edge creation times in Twitter, prove its good theoretical properties and demonstrate explicitly how the error should decrease as a function of follow rate. The method is validated using members of the suggested users list and Twitter celebrities, and we find that is both very accurate and robust to link deletions over time. We use the inferred times to answer many of the above questions through a detailed study of the temporal properties of this Twitter subgraph. Our analysis reveals the importance of the Twitter interface in driving followers to the subgraph and case studies indicate the qualitative magnitude of these effects during different phases of the interface. Examining the distribution of inferred timestamps reveals that more than half of the edges in the subgraph, a non-negligible fraction of the total number of edges in Twitter's social network, formed within one month of the user joining Twitter. Furthermore, cyclical behavior with a period of 24 hours shows that users tend to create edges, and by extension, be logged into Twitter, around the same hour each day as when they created their account. Finally, we demonstrate through several examples that real-world events can correlate strongly with the attractiveness of a celebrity to followers. These results, all temporal in nature, are captured by a single network snapshot in combination with the timestamping method.

Online social networks have attracted much attention as topics for academic study [42, 60].

Many recent papers have demonstrated that online social networks have some of the typical characteristics of real-world networks [66] including short path-lengths, clustering, and heavy-tailed distributions in the number of connections. These distributions are often claimed power-law, although this requires careful study [17].

Twitter, as one of the major online social media websites, has not escaped scrutiny. Communities in Twitter's social network who tweet about similar topics and interests were studied by [34]. Huberman et. al. [32] studied the interaction patterns underlying the social network, suggesting that only a portion of the edges matter for communication over Twitter. Parts of the graph collected under three separate methodologies were analyzed and compared in [41]. More recently, a network analysis based on data collected through breadth-first search was performed by [43] who found a non-power-law follower distribution and low following reciprocity. None of these studies captured the whole of Twitter's social network though, and a discussion of whether network measures are robust under imperfect data is contained in [12].

The interest in online social networks goes well beyond static network analysis. Questions regarding the dynamical evolution of a social network are often very interesting, but also difficult to answer. The dynamical social networks of Flickr and Yahoo! 360 were studied in [42] which had access to precise event times, like those we wish to recover for Twitter. Learning the time intervals in which events occur only from repeated crawling can result in bias for studying certain influence models over social networks [19]. In [51], several networks were shown to densify over time, with the number of edges growing superlinearly with the number of vertices, and average distances shrunk with network size. These novel insights, which contradicted standard views, were not possible without temporal data.

The empirical analysis of mechanisms for network growth (cf. [22]) also requires such data and has occurred at two scales. Macroscopic observations such as that done by [35, 70] found that preferential attachment, a particular mechanism, does appear to hold in certain empirical networks. Microscopic investigations of social networks, at the scale of individual edge placement, has recently been suggested by [48] who compute the likelihood of a host of network formation mechanisms, although not for Twitter. A specific investigation of Twitter was done by [72] who demonstrate the importance of triangle closure in formulating ties. At a smaller scale, triangle closure among many other tie formation mechanisms was investigated in [26]. However, the large-scale study of mechanistic explanations for Twitter's network evolution is limited by the

lack of temporal edge placement data from Twitter. Triangle closure is a special case that can be studied with information from the Twitter API directly. We now show how to gather temporal edge placement data for Twitter and bypass this limitation.

## 3.2 The Algorithm

In this section, we define our timestamping method to infer edge creation times. To understand the procedure, we need to describe the relevant temporal information available from Twitter. There is a single API query that returns the current followers of a particular user in the reverse order in which they followed that user. So even though the time at which the network edges were created is not provided, the order of their creation is known. Another API query can return who a particular user follows, the so-called friends list, and again, this list is returned in reverse temporal order of edge creation. While in general the combination of these local orders is not sufficient to recover a total temporal order of edge placements, combining the friends and follower lists has been useful for studying triangle closure [72].

While the global order of edge placement would be interesting enough to recover, the timestamping method goes beyond this to estimate the actual time at which edges are created. For this feat, we use more than these local orders. Surprisingly, it suffices to just consider each user and their followers individually after incorporating other temporal information from Twitter. Another separate query can map the user identifiers returned by the follower lists to account creation time.

These user creation times along with the edge ordering for a chosen user, will be the input to our procedure. Timestamping a whole collection of users' followers is done through repeated application to each user in turn. Because we apply this method to Twitter's celebrities, for the sake of convenience, we refer to the user chosen for timestamping as a celebrity.

We estimate the edge creation time for any follower of a celebrity by positing that it is equal to the greatest lower bound that can be deduced from the edge orderings and follower creation times for that celebrity. To make this explicit, we define a few relevant variables and then compute this greatest lower bound.

Consider a particular celebrity and let  $U$  be the set of all users following that celebrity. Let  $C_u$  be the creation time of a user  $u \in U$ . Naturally,  $C_u \leq F_u$ , where  $F_u$  is the actual unknown time at which user  $u$  followed the celebrity, for all  $u \in U$ . From the local order contained in this

celebrity’s follower list, we know that  $F_u \leq F_v$  if and only if  $u$  appears before  $v$  in the follower list. These inequalities form the basis of our lower bounds.

Focus on a particular user  $u$  following the celebrity. From the local order, we can construct  $V(u) \subseteq U$  defined to be the set of all users  $v \in U$  such that  $F_v \leq F_u$ . The complement of this set,  $V(u)^c$ , are the users who follow the celebrity after  $u$ . Every user  $v \in V(u)$  (which includes  $u$ ) provides a lower bound on  $F_u$  because  $C_v \leq F_v \leq F_u$ . Users in  $v \in V(u)^c$  for which  $F_v > F_u$  do not provide such a bound because they could have been created before or after the follow time of  $u$ . We take the maximum over all the bounds provided by  $v \in V(u)$  and use that as our estimate for the follow time of  $u$ , denoted  $\hat{F}_u$ , as follows

$$\hat{F}_u = \max_{v \in V(u)} C_v. \quad (3.1)$$

We call any user  $v$  who is the argument of this maximum for user  $u \in U$  a *record-breaker* for user  $u$ . If  $v$  is a record-breaker for  $u \neq v$  then  $v$  is a record-breaker for itself. A simplified definition of record-breaker is thus a user  $u$  that has creation time greater than all preceding users in the follower order. Note that a user is, or is not, a record-breaker for each celebrity that they follow independently.

Our algorithm embodied in Eq. 3.1 is then to identify the record-breakers of the celebrity and assign each follow time to be at the creation time of the most recent record-breaker.

### 3.2.1 Theoretical analysis

In this subsection, we demonstrate that under circumstances appropriate to Twitter’s celebrities, the actual follow times are concentrated about the estimated follow times using the record-breaker users’ creation times. To analyze the inference, we consider a model of following for a given celebrity: Fix creation times  $C_u$  for all users  $u$  that will follow the celebrity. For each user  $u$ , draw an independent, identically distributed non-negative random variable  $L_u$  from an arbitrary latency distribution  $L$  that represents how long  $u$  waits until he decides to follow this celebrity. The probability density function of the latency distribution is given by  $f(t)$ , where  $f(t)$  allows arbitrarily small latencies. So for each user  $u$  the actual follow time is given by  $F_u = C_u + L_u$ .

For simplicity, the creation times are assumed spaced uniformly with time interval  $\lambda$  between each user and the first user is created at time 0. The sequence of creation times is then  $0, \lambda, 2\lambda$ , etc. Let  $P(F_u - \hat{F}_u > \delta)$  be the probability that the error in the inferred following time for user  $u$  is greater than  $\delta$ . (Remember that  $\hat{F}_u \leq F_u$  so the error is always non-negative.) Our main theoretical result, proved in the appendix, shows the following error bound:

**Proposition 1** *Let  $\epsilon > 0$ . If*

$$\left( \int_{\delta/2}^{\infty} f(t) dt \right)^{\delta/(2\lambda)} \leq \epsilon,$$

*then  $P(F_u - \hat{F}_u > \delta) < \epsilon$  for any user  $u$ .*

Note that as  $\lambda$  goes to zero, the theorem is satisfied for any  $\delta$ , implying that the method becomes arbitrarily accurate in this asymptotic limit. This proves that the follow times become concentrated about their greatest lower bounds for small spacing (i.e. high rates of user creation.) Furthermore, the theorem provides an error guarantee for given values of error probability  $\epsilon$  and spacing  $\lambda$ .

It is not essential that the latency distribution be identical between users, that the spacing be given by  $\lambda$ , or that the distribution allow arbitrarily small latencies. Fundamentally, if the rate of new user arrival for a celebrity is high as defined by the proposition, then the error in the inferred follow times will be small. In the next section, we present a thorough validation of the method on empirical Twitter data and demonstrate this negligible error explicitly.

### 3.3 Application of the Algorithm

In this section, we study the celebrity subgraph formed by the 1,508 accurate celebrities. What insights can we now gain that would not be possible without knowing when social links were formed? We first perform a broad analysis of the celebrity subgraph in Section 3.3.1 and then we examine typical accounts in Section 3.3.2. We focus largely on temporal analyses of this subgraph as this is the novel information provided by our method.

#### 3.3.1 Broad analysis of celebrity subgraph

There are 74,184,348, or about 75 million, unique users who follow at least one of the 1508 accurate celebrities. For reference, we estimate the total number of unique users on Twitter to be



around 190 million. So a broad spectrum of user accounts are captured in the subgraph. Some of these unique users are themselves celebrity accounts, so the subgraph is not entirely bipartite. Celebrities do follow each other.

The accurate celebrity subgraph has a total of 835,117,954, or about 835 million, directed edges in it which is actually a non-negligible fraction of edges in Twitter's social graph. A recent study of Twitter as a whole, gathered by breadth-first search, collected 1.47 billion edges in total [43]. An estimate of the total number of edges by the present authors suggests there are around 7 billion edges in the present social graph.

The left window of Figure 3.1 displays the fraction of celebrities with greater than  $k$  followers as a function of  $k$ . The plot is on a log-linear scale and the fairly straight line indicates that the distribution looks exponential. Around 20% of the accurate celebrities have more than a million followers. The right window of Figure 3.1 displays the fraction of users following  $k$  celebrities as a function of  $k$  on a log-log scale. One feature that stands out is the existence of three peaks in the distribution at following 20, 241, and 461 celebrities.

We have been unable to precisely determine the cause of the 241 and 461 peaks, but following 20 celebrities has a simple explanation. It is due to the original formulation of the suggested users list. The suggested users list, in its original design, gave new users the opportunity to automatically follow 20 users randomly selected from a pre-selected collection of users. The default option was to follow all 20 users, but one could click this off to follow a particular subset. The motivation behind the suggested users list was to provide interesting (hand-picked by Twitter) accounts for a new user to follow. According to this article [88], the suggested users list on July 16, 2009 had 241 users on it which is probably the cause of the peak at 241 celebrities. We have been unable to determine if at some time the suggested user list had 461 accounts on it. These peaks constitute prominent evidence that Twitter's interface has dramatically effected the celebrity subgraph.

Further indications can be seen in Figure 3.2 where the blue curve shows the number of edges created in the accurate celebrity subgraph per hour as a function of time. We have labeled three distinct changes in this total celebrity follow rate.

These changes correspond to three distinct adjustments to Twitter's user interface. The first label (1) is the introduction of the suggested users list which occurred around February 2009 [81, 83]. Using the account creation times of the users who follow 20 celebrities suggests

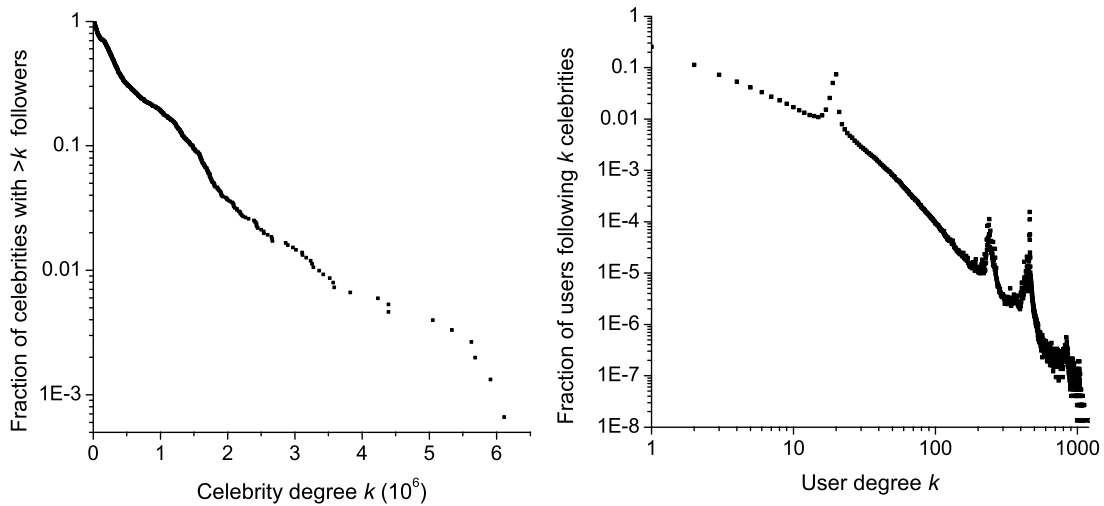


Figure 3.1: Left side: The complementary cumulative distribution function for the number of followers of a celebrity. Note that this is a log-linear scale. Right side: The distribution of the number of celebrities followed by a user plotted on a log-log scale. Notice the three peaks at  $k = 20, 241$  and  $461$ .

that the actual date was Feb. 13, 2009, when there was a large upward surge in following 20 celebrities. Label (2) shows when the old suggested user list was changed to its current format on Jan. 21, 2010 [84] at which point the number of followers drops dramatically. The updated format displays a number of categories such as science and entertainment and new users are encouraged to follow suggested users corresponding to their interests.<sup>1</sup> Much of the drop in volume that occurs on Jan. 21, 2010 is due to the suggested users list no longer defaulting to follow 20 celebrities. Correspondingly, there is a sharp decline in the number of users following 20 celebrities after Jan. 21, 2010.

The last change (3) is due to the introduction of the “users you may be interested in” (or “Suggestions for You”) feature which was rolled out on July 30, 2010 [82]. This feature suggests accounts to existing Twitter users that they might want to follow. We see another upsurge in celebrity follow rate around the same time.

One possible explanation for these rapid changes is that the introduction of a feature, or change in user recommendation system, by Twitter adjusts the rate at which accounts are created. We test this hypothesis by computing the rate at which accounts were created for Twitter, shown in the green curve of Figure 3.2. While there is perhaps a slightly contemporaneous increase in

<sup>1</sup>The suggested users list could also be reached from the Twitter homepage in both of its implementations.

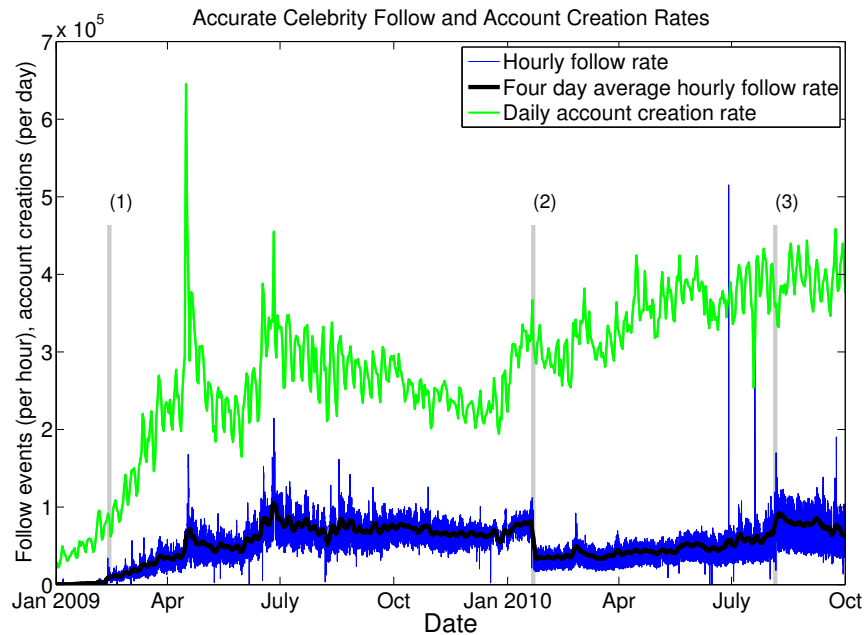


Figure 3.2: The total celebrity follow rate (follow events per hour) and Twitter account creation rate (accounts created per day) over time. The three labels correspond to the introduction of the suggested users list, the update to the suggested users list, and introduction of “users you may be interested in”. The black smoothed curve shows a four day average of the celebrity follow rate.

total celebrity follow rate and account creation when the old suggested user list is introduced, the increase in user creation is not sustained. Similarly, the change in follow rate due to the switch from old to categorical suggested user list and introduction of “users you may be interested in” is not explained by changes in account creation. Since the creation rate of Twitter accounts is unable to account for the changes in celebrity follower rate due to altered Twitter features, the more plausible explanation is instead that these features altered how users discover and follow celebrity accounts.

In order to analyze these effects further, we examine several typical accurate celebrities on the suggested users list as case studies in the next section.

### 3.3.2 Impact of the Suggested Users List

Given that the overall celebrity follow rate halved when Twitter switched to the categorical suggested users list, it is clear that being on the suggested users list increases the acquisition of new followers substantially. Anil Dash, a tech blogger and entrepreneur, has written about his experiences being on the old version of the suggested users list [8] and is an illustrative example.

At the time of our data collection, Mr. Dash had 332699 followers in total. In figure 3.3, we show the fraction of Mr. Dash’s follow events per day using the inferred timestamps.

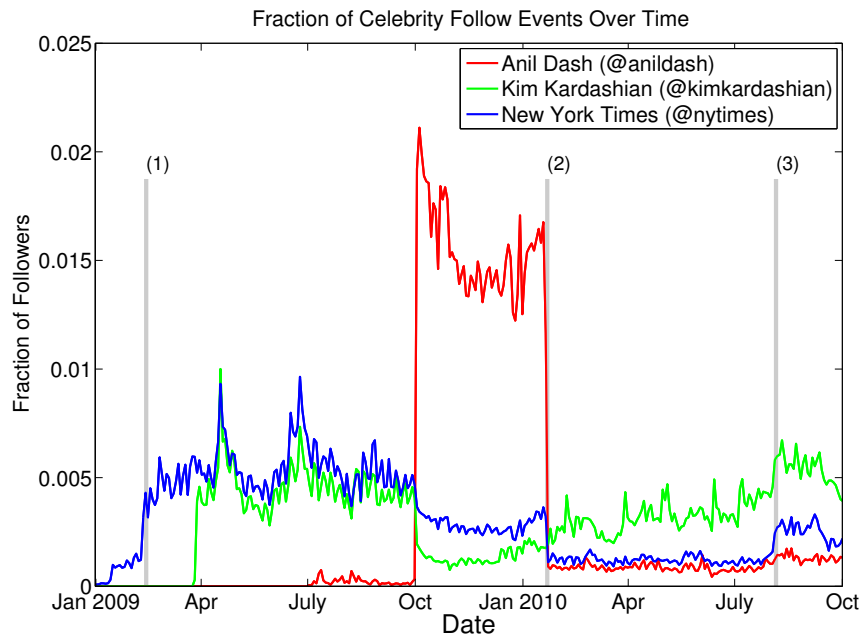


Figure 3.3: The fraction of follow events for each celebrity per day as a function of time. The three labeled grey lines are the times of the interface changes described in Sec. 3.3.1.

Very shortly after being put on the old suggested user list on Oct. 2, 2009, Mr. Dash’s rate of gaining followers increased greatly. During his time on the old suggested user list, he gained around 2,500 new followers per day compared to his previous average of about 50 per day. When Twitter transitions to the categorized suggested user list, his following rate drops significantly to around 100 followers per day. Interestingly, this is still higher than before his presence on the old suggested user list. We consider two possible explanations for this continued popularity. Many models of network formation assume that edges are “sticky” in the sense that gaining followers increases the rate at which you will gain followers in the future. It is reasonable that the large number of followers gained from being on the old suggested users list had this effect for Mr. Dash. Alternatively, his account could have been present immediately in the categorized suggested users list and this mechanism could account for the additional followers. Mr. Dash is (as of October 20, 2010) in the technology category of the suggested users list, but as the list changes over time, we cannot say if he was on the list in January. A smaller, but still evident, increase in follower rate to around 200 followers per day on average occurs during the

introduction of the “users you may be interested in” feature. This increase is not nearly the boost given by the old suggested users list, but it is certainly non-negligible.

Also shown on the figure are the corresponding curves for the New York Times and Kim Kardashian. The New York Times account was created before the old suggested users list and immediately benefits from its introduction at label (1). Kim Kardashian apparently was placed onto the list shortly after her account was created as her curve tracks the New York Times fairly closely during the time of the old suggested users list. In October, when Mr. Dash is placed onto the suggested user list, both @nytimes and @kimkardashian drop in their follow rate. It could be that the suggested users list expanded (perhaps to 461 from 261 accounts) or they were removed from the suggested users list. Judging by the sharp decline in @nytimes fraction at (2), it was likely on the suggested users list with Mr. Dash. Then finally the introduction of “users you may be interested in” benefited @nytimes and @kimkardashian, although again not as much as the old suggested users list. These case studies illustrate that a wide range of different Twitter celebrities experienced similar follow behavior due to the interface.

Besides knowing when edges are created, we are also interested in how long users wait to follow celebrities after they join Twitter.

### **3.3.3 Measuring following latency**

In our theoretical analysis, users’ following behavior is determined by a latency distribution. We examine the actual latency of users, the differences between their account creation time and following time. Because our data only contains users who have followed the celebrities when the network snapshot is taken, early users may exist who will follow the celebrities in the future and have long latencies. Ignoring these users, and their long latencies, would bias any attempt to empirically determine the latency distribution, especially because we cannot identify which users will ever decide to follow a celebrity.

So instead we measure the conditional probability that a user waits  $t$  seconds to follow the celebrity given that they follow the celebrity within a month of account creation. In Figure 3.4, this unnormalized probability is estimated by the number of follow events derived from users created more than a month ago on a log-linear scale in hourly bins. The large concentration at zero latency is caused by the set of record-breaker users. Of those users who follow within a month, 86 percent follow within 24 hours and 90 percent follow within six days. If a user is

going to follow a celebrity within a month of joining Twitter, they are most likely to do so nearly immediately after joining.

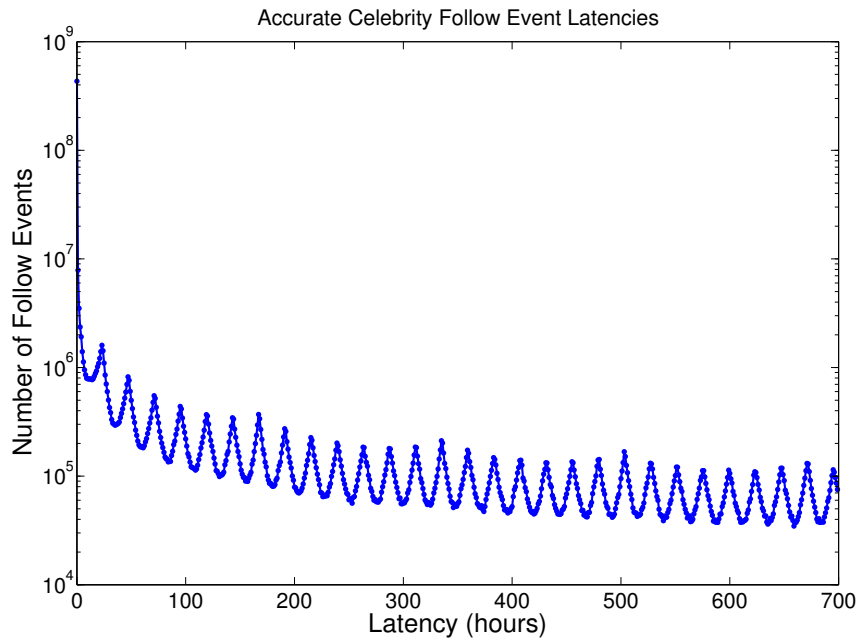


Figure 3.4: The number of follow events binned by hour as a function of latency for the follow events of users created before September 1, 2010.

The intriguing periodicity in the distribution occurs over 24 hour intervals and this could be because users prefer to follow celebrities around the same hour of the day that they created their account. One could imagine that a Twitter user logs onto Twitter around the same time everyday. We check this interpretation in Figure 3.5 which is a heatmap on a log-scale showing the number of follow events created during the hour on the y-axis for users created during the hour on the x-axis. We only include latencies greater than a day to eliminate the large contribution due to the record-breakers. This figure is consistent with our interpretation of the latency distribution as it is nearly diagonal. Moreover, the peak along the diagonal indicate that 4-10 pm EST is a popular time to both follow celebrities and create accounts.

The fraction of each celebrity’s followers who followed the celebrity within a month of joining Twitter varies widely over the celebrities with an average of 65% and a standard deviation of 18%. This large fraction of each celebrity’s followers translates into nearly 580 out of the 835 million edges with latency less than a month. If we change the scale from a month to a day, on average 48% of a celebrity’s followers followed them within a day. Again translated into edges,

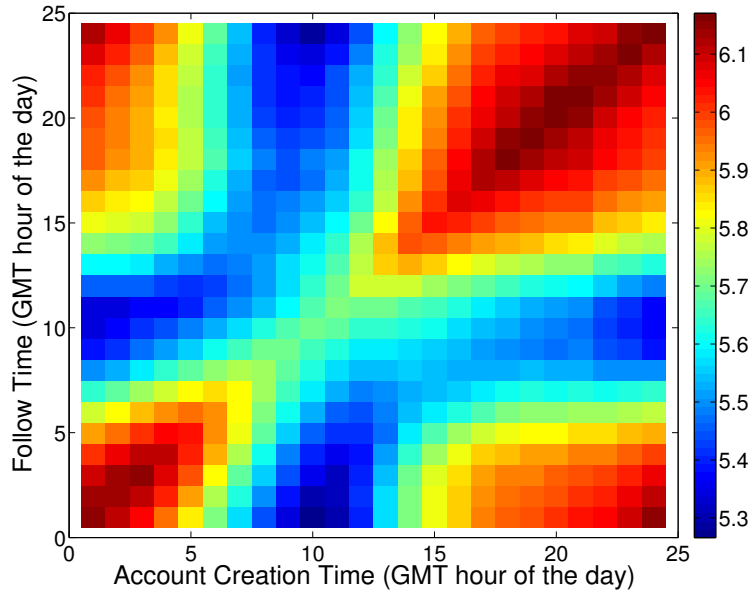


Figure 3.5: A heatmap of the creation time versus follow time over all celebrities with latencies greater than one day on a log-scale. The hours represent the GMT timezone.

about 451 million edges have latency less than a day. In fact, about 140 million of the edges are due to record-breakers and hence are given a latency of zero. While old users do follow celebrities with occasionally large latency, low latency edges are dominant.

### 3.3.4 Celebrity popularity and real-world events

We have seen that the rate at which accurate celebrities gain followers is plausibly changed by adjustments to Twitter’s interface. In this section, we examine whether the rate at which a celebrity receives followers could also plausibly be changed by real-world events.

For our first demonstration of a plausible real-world event that changed Twitter, during the Iran election in late June 2009, Twitter became a vehicle of communication among Iranian internet users planning protests and rallies. Twitter was popularized by the mainstream media at this time, and we witness a sharp increase in the number of new accounts in July 2009. Conveniently for our purposes, celebrities often show up in the national news for particular events such as political rallies, concerts, or sporting events. Are such single day events important to the temporal evolution of Twitter’s celebrity follower subgraph?

It is not effective to analyze absolute follow rates to answer this question because the abso-

lute rate depends on the total rate of user account creation which varies substantially as shown in Figure 3.2 with occasionally sharp changes. To compensate for such overall variation, we consider whether the relative rate, which we call relative popularity, of a celebrity changes due to real-world events.

The relative popularity  $f_i(t)$  is an estimate of the probability that a user who follows a celebrity at time  $t$  decides to follow celebrity  $i$ . This relative popularity is normalized so that  $\sum_i f_i(t) = 1$ , where the sum is over all celebrities and the relative popularity is zero for a nonexistent celebrity at time  $t$ . We compute it using the following sliding window:

$$f_i(t) = \frac{|\text{Connections to } i \text{ within } t - \Delta \text{ and } t + \Delta|}{|\text{Edges created within } t - \Delta \text{ and } t + \Delta|}, \quad (3.2)$$

where the variation of  $f_i(t)$  is assumed to be at a longer time-scale than window width  $\Delta$ . We checked several values  $\Delta$  to ensure consistent results and decided to use a window width equal to a week with  $t$  samples spaced per day. A useful comparison is the relative popularity if followers were placed randomly, which is simply  $1/n(t)$  where  $n(t)$  is the number of celebrities that exist at time  $t$ .

We computed these curves utilizing the top 50 celebrities and in Figure 3.6, we display the resulting relative popularity values for the top 10 celebrities. These values are clearly varying over time, and are far from the predictions of random attachment represented by the black line. The behavior of the relative popularity when a new celebrity joins Twitter differs widely. Oprah Winfrey and Ellen DeGeneres, for example, have a quick spike upwards in relative popularity, but Justin Bieber, one of the most popular of the top 50 currently, begins with a small relative popularity that gradually increases over time. The relative popularity shows large variations, including several prominent peaks and drops that are not due to Twitter's interface. One such drop is near June 25, 2010 (arrow 5) where the rapper Soulja Boy, not on the top 10, gained roughly half a million followers over a few days, garnering a relative popularity value of nearly twenty-five percent. A search of blog posts and news articles reveals that Soulja Boy deleted his Twitter account called @SouljaBoyTellEm and switched to an account called @SouljaBoy. One explanation is that these users followed Soulja Boy from his previous popular account. Alternatively, the hashtag #IfSouljaBoyWasARapper was a trending topic, which means that tweets containing the phrase #IfSouljaBoyWasARapper were extremely popular on Twitter around June 25. While the humor was decidedly unfavorable to Soulja Boy, these tweets may have had a positive effect



on his relative popularity.

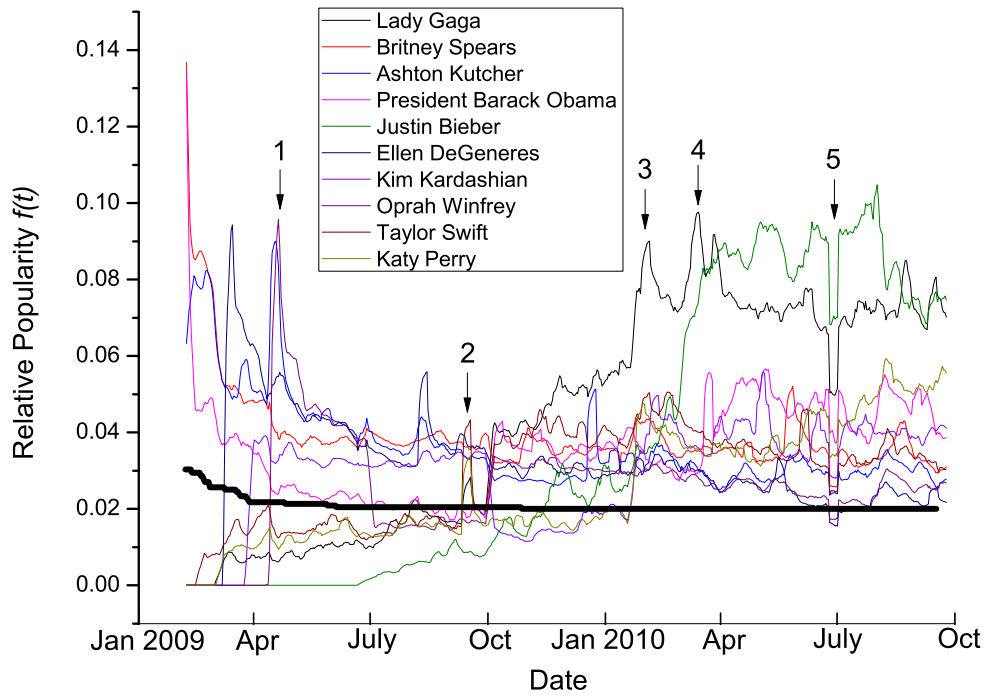


Figure 3.6: The relative popularity as a function of time for the top 10 celebrities. The random attachment prediction is shown in bold. Labeled arrows correspond to events discussed in the text.

For many other cases, we can also identify spikes in relative popularity as corresponding to real-world events that plausibly explain increased Twitter popularity. For example, Lady Gaga performed at the Emmy’s on Feb. 1st, 2010 (arrow 3) and released her music video “Telephone” on March 13, 2010 (arrow 4). Even more interestingly, the peaks that occur simultaneously for several celebrities appear to be due to events involving them together. On Friday April 17, 2009 (arrow 1) Ashton Kutcher, who just succeeded in reaching one million Twitter followers before CNN Breaking News, appeared on Oprah’s TV show, during which she joined Twitter [68]. They both received large boosts in relative popularity from this event and we suspect that Ashton and Oprah are collectively responsible for the largest gain in Twitter accounts ever that occurred on this day in April (see Figure 3.2). Lady Gaga also performed at the MTV Video Music Awards on Sept. 12, 2009 (arrow 2) along with Taylor Swift and Katy Perry. [63] All three of them show an increase in relative popularity at this time. Unfortunately, Kanye West, who was involved in an infamous incident with Taylor Swift that evening, was not on Twitter at the time.

### 3.4 Proof of proposition

Proof: Pick an arbitrary user  $u$  to compute the error probability. We start from

$$P(F_u - \hat{F}_u > \delta) = \int_0^\infty P(F_u - \hat{F}_u > \delta | L_u = t) f(t) dt.$$

Since  $\hat{F}_u \geq C_u$ , the error is at most equal to  $t$  for fixed  $L_u = t$ , we can change the bottom limit of integration to  $\delta$  to get

$$P(F_u - \hat{F}_u > \delta) = \int_\delta^\infty P(F_u - \hat{F}_u > \delta | L_u = t) f(t) dt.$$

Consider a fixed value  $t$  and define  $N_1(u)$  as the set of users  $v$  such that  $F_u - \delta/2 < C_v < F_u$  and  $N_2(u)$  to be the set of users  $v$  such that  $F_u - \delta \leq C_v \leq F_u - \delta/2$ . The probability  $P(F_u - \hat{F}_u > \delta | L_u = t)$  is equal to the probability that all these users have  $F_v > F_u$ . If that happens, all of these users are in  $V(u)^c$  and therefore,  $\hat{F}_u = \max_{v \in V(u)} C_v < F_u - \delta$ . The condition  $F_v > F_u$  for fixed  $L_u = t$  is met if  $L_v > C_u - C_v + t$ . Let  $Q_{uv}(t) = C_u - C_v + t$  and  $P(L > x) = \int_x^\infty f(t) dt$ . Then we can express the conditional error probability as

$$P(F_u - \hat{F}_u > \delta | L_u = t) = \prod_{v \in N_1(u) \cup N_2(u)} P(L > Q_{uv}(t)).$$

We upper-bound this expression as follows:

$$P(F_u - \hat{F}_u > \delta | L_u = t) \leq \prod_{v \in N_2(u)} P(L > Q_{uv}(t)).$$

Note that  $Q_{uv}(t) \geq \delta/2$  for  $v \in N_2(u)$ . Then

$$\begin{aligned} P(F_u - \hat{F}_u > \delta | L_u = t) &\leq P(L > \delta/2)^{|N_2(u)|} \\ &= P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor}. \end{aligned} \tag{3.3}$$

This bound no longer depends on  $t$ . So

$$\begin{aligned} P(F_u - \hat{F}_u > \delta) &\leq \int_{\delta}^{\infty} P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor} f(t) dt & (3.4) \\ &< P(L > \delta/2)^{\lfloor \delta/(2\lambda) \rfloor + 1} \\ &< P(L > \delta/2)^{\delta/(2\lambda)}, \end{aligned}$$

which completes our proof.



# Chapter 4

## User Interactions in Twitter

### 4.1 Introduction

As social media applications gain richer sets of features, they come to contain increasingly diverse connections among their users. The microblogging site Twitter is one example in which users share links to content, pass on messages, and initiate messages with an intended target. These directed messages, which we call *@-messages*, signal a communication link between two users and can represent many different forms of interaction. Indeed, earlier research has shown that Twitter contains a large amount of social activity between users who interact with each other as peers, as well as a large amount of information-seeking and information-sharing activity in which users interact with celebrities, news sources, and other types of high-visibility accounts [44, 73].

A challenge when studying an environment such as Twitter is that these types of connections are superimposed in a single communication network. Therefore, it is important to develop techniques capable of classifying the links in the underlying network according to the different activities that they represent. To this end, we formulate the problem of predicting *link reciprocity*, and develop a set of techniques for this purpose.

Reciprocity captures a basic way in which different forms of interaction on a site like Twitter take place. When two users  $v$  and  $w$  interact as peers, one expects that *@-messages* will be exchanged between them, passing in both directions — we consider this to be a *symmetric*, or *reciprocated*, interaction. On the other hand, if a user  $v$  sends multiple messages to a celebrity

<sup>0</sup>This chapter is adapted from Cheng, Romero, Meeder, Kleinberg [16].

or news source  $w$ , it is likely that  $w$  will not send messages in return — this is an *asymmetric*, or *unreciprocated* interaction. Which features characterize the difference between reciprocated and unreciprocated relationships? Can we tell them apart based on properties of the users involved, and properties of their network neighborhoods? Do the sub-networks of reciprocated and unreciprocated links have different structural properties? These are some of the questions we address in this paper; our approach also addresses the broader issue of isolating the different forms of interaction that take place on a complex social media site such as Twitter.

### 4.1.1 Summary of Results

We pursue two main approaches to analyzing reciprocity. The first is to study the problem of reciprocity prediction: we formulate several variants of the problem, all of them oriented around determining whether a link between users  $v$  and  $w$  is reciprocated or unreciprocated; and we identify a set of features based on the characteristics of  $v$ ,  $w$ , and the nodes connected to  $v$  or  $w$ . (The precise definitions for these variants of the problem will be given in the next section.) Our analysis extracts features that have strong predictive power for this task. We find that differences in reciprocity can be related to the notion of *status*. Roughly speaking, people with similar status often participate in reciprocated interactions (e.g. messages between friends), while those with disparate status often participate in unreciprocated interactions (e.g. messages from fans to celebrities). In particular, we find that measures that formalize the relative “flow” of links from  $v$  to  $w$ , compared with the corresponding flow from  $w$  to  $v$ , constitute an important source of information for this task.

Our second approach involves comparing the structure of two subgraphs, one consisting of just the reciprocated links and the other consisting of just the unreciprocated links. We find that these structures exhibit important differences, including the presence of greater clustering and a larger giant component in the subgraph of symmetric links. Moreover, we find that almost all highly active Twitter users take part in at least some reciprocated interactions, while a non-trivial fraction take part in no unreciprocated interactions.

### 4.1.2 Related work

As noted above, several recent papers have discussed the heterogeneity of relationship types on Twitter, although they do not consider reciprocity as a measure [44, 73]. In the context of email, Tyler and Tang considered the dynamics of replying to messages, which is the analogue of reciprocation in that domain [86]; however, the specifics of their analysis are quite different from what we pursue here.

Predicting reciprocity is related to other prediction tasks concerned with the links of an underlying network, but it is different in several important respects. Specifically, the *link prediction problem* seeks to identify links  $(v, w)$  that are currently missing in a network snapshot but are likely to form in the near future [53]. A key contrast between reciprocity prediction and link prediction is that the formation of any particular link is a rare event, whereas reciprocating an existing link  $(v, w)$  with the reverse edge  $(w, v)$  is common on sites such as Twitter. For this and other reasons, we find that features that have been observed to work well for link prediction are not the most effective for reciprocity prediction.

There has also been recent research predicting the *strengths* [23] and the *signs* [49] of links in on-line social networks. These works form interesting contrasts with reciprocity prediction. In particular, it is possible for a strong directed tie to be either reciprocated or unreciprocated. For example, an avid follower of the New York Times might regularly generate messages such as “@nytimes reports that ... ” without ever receiving a message from the @nytimes Twitter account. In the problem of sign prediction the notion of status is crucial [49], as it is in our work; but there are important distinctions in that reciprocated and unreciprocated links can easily exhibit either type of sign.

## 4.2 Problem definition

We now formalize the problem of predicting link reciprocity. The communication network is represented as a directed graph  $G = (V, E)$ , with an edge  $(v, w)$  indicating that  $v$  has sent  $w$  at least one @-message. (In keeping with the focus on communication activities, we use @-messages to define all the networks in our analysis, rather than other relationships such as  $v$  following  $w$ 's account.) The input to our prediction problem is the graph  $G$  and a node pair  $\{v, w\}$ , where at least one of the edges  $(v, w)$  or  $(w, v)$  is present in  $G$ ; information about all edges of  $G$  is pro-

vided, except that the presence or absence of the two potential edges  $(v, w)$  and  $(w, v)$  has been hidden. Our task is to predict the direction of edges between  $v$  and  $w$ , and we note that this can be formalized in two distinct ways. First, we consider a formulation in which we decide whether a  $\{v, w\}$  relationship is symmetric (that is, both  $(v, w)$  and  $(w, v)$  exist) or is asymmetric (only one of the directed  $(v, w)$ ,  $(w, v)$  relationship exists). In the second formulation, we ask whether the  $(w, v)$  edge is present given that the  $(v, w)$  edge exists. Intuitively, predicting a symmetric relationship between nodes is a more difficult task than predicting reciprocation in a specific direction, and we show that this is indeed the case.

### 4.2.1 Notation

The number of messages produced by users on Twitter exhibits a long-tail distribution – many users produce only a small number of messages. In this work, we focus on users who have produced a large number of @-messages, so that we are studying a user population for whom Twitter is a significant communication medium. We consider subgraphs of the form  $G_n = (V_n, E_n)$ , where  $V_n = \{v \mid v \in V, v \text{ sent } \geq n \text{ @-messages}\}$  and  $E_n = \{e = (v, w) \mid e \in E, v \text{ and } w \in V_n\}$ . This subgraph thus captures the @-messaging interactions between individuals who are prolific Twitter users. We use the notation  $v \xrightarrow{k} w$  to indicate that  $v$  sent at least  $k$  @-messages to  $w$ , and  $v \xrightarrow{=k} w$  to indicate that  $v$  sent exactly  $k$  @-messages to  $w$ . From this definition we can parametrize reciprocity in terms of  $k$ . We say that an edge  $(v, w)$  is *reciprocated* if both  $v \xrightarrow{k} w$  and  $w \xrightarrow{k} v$ , and is *unreciprocated* if  $v \xrightarrow{k} w$  and  $w \xrightarrow{=0} v$ . Let the set of reciprocated edges be denoted  $E_k^r$  and the set of unreciprocated edges be denoted  $E_k^u$ . Finally, let  $\text{deg}^-(v)$  and  $\text{deg}^+(v)$  respectively denote the indegree and outdegree of node  $v$ ,  $\text{msg}^-(v)$  and  $\text{msg}^+(v)$  be the numbers of messages received and sent by  $v$ ,  $\Gamma^-(v) = \{w \mid (w, v) \in E\}$  be the set of people who sent messages to  $v$ , and  $\Gamma^+(v) = \{w \mid (v, w) \in E\}$  be the set of people to whom  $v$  sent messages.

### 4.2.2 Dataset description

We extracted the @-message graph from a large crawl of Twitter that took place between August 2009 and January 2010. More than three-billion messages from over 60 million users were collected in this data set [76]. The @-message graph is constructed by looking at messages a user  $v$  authors which mention user  $w$  at the beginning of the tweet. The graph  $G$  of users



who authored at least one @-message contains 12,795,683 distinct users who sent a total of 819,305,776 @-messages, with 156,868,257 distinct directed interactions. We focus our analysis on the subgraph  $G_{1000}$  induced by users who authored at least 1000 @-messages. In  $G_{1000}$ , which includes 181,033 users, we find that  $|E_{10}^r| = 797,342$  and  $|E_{10}^u| = 349,258$

### 4.3 Methods for Reciprocity Prediction

Intuitively, features that capture whether  $v$  and  $w$  have similar status or a similar social circle should be potentially useful in predicting reciprocation. This section presents the various features that we use for predicting reciprocity in networks. Each feature corresponds to a method that assigns a value  $\text{val}(v, w)$  to a node pair  $(v, w)$ , or a value  $\text{val}(v)$  to a single node  $v$ . For each feature, we look at its value and whether the edge in question is reciprocated; this data is used for training models to predict reciprocity.

Given the values corresponding to all node pairs (or nodes) in question, we can then choose threshold values or ranges where we predict reciprocity, and predict a lack thereof in the complementary region. We consider a simple threshold classification scheme which predicts that a node pair  $(u, v)$  is unreciprocated if the feature value is less than (or greater than) some threshold, and is reciprocated otherwise. For each feature, we determine the threshold value  $\text{val}_{OPT}$  and threshold direction (less than / greater than) to maximize prediction accuracy according to this threshold classifier. For example, a sufficiently high number of mutual neighbors for the nodes  $v$  and  $w$  could strongly indicate the existence of a reciprocated link between them. The list of features we consider is summarized in Table 4.1.

As previously mentioned, there are two ways that we can formulate the prediction problem, and we present four different formulations. The first addresses the question of symmetry, while the other three examine the problem of reciprocity in which the available information about the nodes in question is limited:

1. SYM (predicting symmetry): predict whether both  $(v, w)$  and  $(w, v)$  exist, or whether exactly one of  $(v, w)$  or  $(w, v)$  exists, using information about  $v$  and  $w$  but not about the presence or absence of communication between them.
2. REV (predicting a reverse edge): predict whether a reverse edge exists given that the forward edge  $(v, w)$  exists, using information about  $v$  and  $w$ .

3. REV- $w$  (predicting a reverse edge using only  $w$ ): predict whether a reverse edge exists given that  $(v, w)$  exists, but using only information about  $w$ .
4. REV- $v$  (predicting a reverse edge using only  $v$ ): predict whether a reverse edge exists given that  $(v, w)$  exists, but using only information about  $v$ .

Within this framework, we can compare the predictive power of specific features of the @-message graph, as well as more complex classifiers (such as decision trees) that utilize multiple features.

### 4.3.1 Degree and message features

It seems intuitive that the relative indegree or outdegree of nodes would indicate whether a pair of nodes are in a one-sided or two-sided relationship. If both have a similar indegree, this might indicate that they have similar social status in the network. In contrast, disproportionate indegrees could indicate that one user is a celebrity and the other is a non-celebrity, making it less likely that their relationship is reciprocated.

We now describe the features in our analysis that are based on degree and message counts. Both relative (e.g., the ratio of indegrees) and absolute (e.g.,  $\text{deg}^-(w)$ ) feature measures are considered:

- *Indegree and outdegree ratio* both measure the ratio of outdegrees or indegrees of two nodes, and we define  $\text{val}(v, w) = \text{deg}^-(v)/\text{deg}^-(w)$  or  $\text{deg}^+(v)/\text{deg}^+(w)$ , respectively.
- *Incoming message and outgoing message ratio* are similar, but uses the total number of messages that a node receives or sends, regardless of the nodes to which messages are sent or from which messages are received. Specifically, we use the analogous measures discussed above for degree, but with  $\text{msg}^-(v)$  and  $\text{msg}^+(v)$  playing the roles of  $\text{deg}^-(v)$  and  $\text{deg}^+(v)$ , respectively.
- *Incoming message/indegree ratio and outgoing message/outdegree ratio* compares the ratio of two nodes' incoming message to indegree ratio or outgoing message to outdegree ratio. A high incoming message to indegree ratio might characterize users who have a small group of friends with which they exchange many messages. Alternatively, a low incoming message to indegree ratio could characterize highly connected people in a network, since the messages they receive are distributed over many more users.

- *Outdegree/indegree ratio* is a heuristic that characterizes the messaging activity of a single node. A large outdegree/indegree ratio might indicate a user of celebrity status because she receives many messages from many followers but sends relatively few messages. Given a pair of nodes we can compute the outdegree/indegree ratio as  $\text{val}(v, w) = \frac{\text{deg}^+(v)}{\text{deg}^-(v)} / \frac{\text{deg}^+(w)}{\text{deg}^-(w)}$ .

### 4.3.2 Link prediction features

It is not intuitive whether methods that work well for link prediction would work well in predicting reciprocity. While link prediction asks whether an edge between two nodes exists, reciprocity prediction asks whether a pair known to have at least a directed edge in fact has a bi-directional pair of edges. The following are some measures used for link prediction:

*Newman* [65] showed that the number of common neighbors in a collaboration network can be a predictor of future links.

*Mutual neighbors* calculates the number of people to whom both  $v$  and  $w$  send messages ( $|\Gamma^+(v) \cap \Gamma^+(w)|$ ), or the number of people from whom both  $v$  and  $w$  receive messages ( $|\Gamma^-(v) \cap \Gamma^-(w)|$ ).

*Jaccard's coefficient* [77] is a similarity measure that we apply to the concept of mutual neighbors. We calculate the similarity between two sets by taking the ratio of the cardinality of their intersection and their union:

$$\text{val}(v, w) = \frac{|\Gamma^-(v) \cap \Gamma^-(w)|}{|\Gamma^-(v) \cup \Gamma^-(w)|}$$

*Adamic and Adar* [2] defined the similarity between Web sites  $v, w$  to be

$$\sum_{\{x|v,w \text{ share feature } x\}} \frac{1}{\log \text{frequency}(x)},$$

and we similarly define  $\text{val}(v, w)$  to be

$$\sum_{\{x|x \in \Gamma^-(v) \cap \Gamma^-(w)\}} \frac{1}{\log \text{deg}^-(x)}.$$

*Preferential attachment* is another popular heuristic in modeling network growth, where the probability that an edge forms with a specific node is proportional to its existing indegree. New-

man [65] and Barabasi et al. [11] showed that the product of the in-degrees of two nodes in a co-authorship network can be a predictor of a future link between the nodes. We apply preferential attachment in a slightly different way and define  $\text{val}(v, w) = \text{deg}^-(v) \cdot \text{deg}^+(w)$  or  $\text{deg}^+(v) \cdot \text{deg}^-(w)$ . Notice that taking the ratio of these two values is equivalent to the outdegree/indegree ratio between two nodes.

*Two-step paths (ratio)* is a simplification of Katz’s [39] measure of status by calculating the number of paths between two nodes. In this work, we only consider paths of length 2, and define  $\text{val}(v, w) = |\text{paths}^2(v, w)|$ , where  $\text{paths}^2(v, w)$  is the set of paths from  $v$  to  $w$  of length 2. The two-step paths ratio is simply the ratio of the number of directed two-step paths from  $v$  to  $w$  to that from  $w$  to  $v$ .

### 4.3.3 Different sets of features

The features above have been shown to work well for the related but different task of predicting the existence of a link. Since our task is to predict whether a link is reciprocated we examine several additional features. For convenience, we further break them down into four sets:

1. Absolute degree/message features - degree, messages, message-degrees, outdegree-indegrees;
2. Relative degree/message features - degree ratios, message ratios, message-degree ratios, and outdegree-indegree ratios;
3. Two-step hop features - mutual neighbors (in and out), and two step paths ( $v$  to  $w$  and  $w$  to  $v$ ); and
4. Link prediction features - all other link prediction features not mentioned

### 4.3.4 Two-step paths

The importance of “friends of friends,” or people two links away from a given node, lends itself to exploring features that directly arise from the directed @-message graph. There are essentially four types of two-step hops corresponding to either the number of common in-neighbors or out-neighbors (mutual neighbors), or the number of directed paths from  $v$  to  $w$  or from  $w$  to  $v$  (two-step paths).

If both  $v$  and  $w$  send messages to many common people, it is likely that they are in the same social circle, or that they mention the same celebrities. If  $v$  and  $w$  receive many messages from

the same group of people, it could be that both  $v$  and  $w$  are in the same community, or that they are celebrities with overlapping fan-bases.

As the number of paths from  $v$  to  $w$  increases, there are two conflicting forces:  $v$  has a stronger source of connections to  $w$ , but at the same time  $w$  is more popular and hence less likely to reciprocate the  $(v, w)$  edge. The reverse case is simpler—intuitively, as the number of paths from  $w$  to  $v$  increases, the likelihood that  $w$  will communicate with  $v$  grows.

## 4.4 Results and Discussion

### 4.4.1 Individual properties

To calculate the accuracy of the individual heuristics, we calculated  $\text{val}$  for each feature on the subset  $E_{10}^r \cup E_{10}^u$  of the graph  $G_{1000}$ , where equal numbers of edges were taken from the sets of reciprocated and unreciprocated edges. This gives a baseline accuracy of 0.500, achievable by predicting that all edges are of one type. We applied the SYM and REV mechanisms to feature sets 2-4, and REV- $v$  and REV- $w$  to set 1.

As described earlier, we picked a threshold value  $\text{val}_{OPT}$  to optimize prediction accuracy: we predicted reciprocity above the threshold, and non-reciprocity below (or vice versa depending on which performed better). Tables 4.3 and 4.4 summarize the performance of each heuristic on the subgraph  $G_{1000}$ ,  $k = 10$ , while table 4.2 summarizes the different mechanisms of prediction for a single heuristic.

In tables 4.3 and 4.4, a star (\*) indicates that reciprocity was predicted when  $\text{val}$  was below the threshold, and a lack thereof indicates reciprocity was predicted when  $\text{val}$  was above the threshold.

In table 4.2, SYM<sup>+</sup> refers to the prediction mechanism where we aim to predict symmetry and predict all edges with values *above*  $\text{val}_{OPT}$  to be reciprocated, and REV<sup>-</sup> refers to the mechanism where we aim to predict whether a reverse edge  $(w, v)$  exists given  $(v, w)$  and predict all edges with values *below*  $\text{val}_{OPT}$  to be reciprocated.

Table 4.1: Reciprocity Prediction Features

Feature	$\text{val}(v)$ or $\text{val}(v, w)$
<i>Absolute degree/message features</i>	
Indegree or outdegree	$\text{deg}^-(v)$ or $\text{deg}^+(v)$
Incoming or outgoing messages	$\text{msg}^-(v)$ or $\text{msg}^+(v)$
Message-degree (in or out)	$\frac{\text{msg}^-(v)}{\text{deg}^-(v)}$ or $\frac{\text{msg}^+(v)}{\text{deg}^+(v)}$
Outdegree-indegree	$\frac{\text{deg}^+(v)}{\text{deg}^-(v)}$
<i>Relative degree/message features</i>	
Indegree ratio	$\text{deg}^-(v) / \text{deg}^-(w)$
Outdegree ratio	$\text{deg}^+(v) / \text{deg}^+(w)$
Incoming message ratio	$\text{msg}^-(v) / \text{msg}^-(w)$
Outgoing message ratio	$\text{msg}^+(v) / \text{msg}^+(w)$
Message-degree ratio (in)	$\frac{\text{msg}^-(v)}{\text{deg}^-(v)} / \frac{\text{msg}^-(w)}{\text{deg}^-(w)}$
Message-degree ratio (out)	$\frac{\text{msg}^+(v)}{\text{deg}^+(v)} / \frac{\text{msg}^+(w)}{\text{deg}^+(w)}$
Outdegree-indegree ratio	$\frac{\text{deg}^+(v)}{\text{deg}^-(v)} / \frac{\text{deg}^+(w)}{\text{deg}^-(w)}$
<i>Link prediction features</i>	
Mutual neighbors (in)	$ \Gamma^-(v) \cap \Gamma^-(w) $
Mutual neighbors (out)	$ \Gamma^+(v) \cap \Gamma^+(w) $
Jaccard's coefficient (in)	$\frac{ \Gamma^-(v) \cap \Gamma^-(w) }{ \Gamma^-(v) \cup \Gamma^-(w) }$
Jaccard's coefficient (out)	$\frac{ \Gamma^+(v) \cap \Gamma^+(w) }{ \Gamma^+(v) \cup \Gamma^+(w) }$
Adamic/Adar	$\sum_{\{x x \in \Gamma^-(v) \cap \Gamma^-(w)\}} \frac{1}{\log \text{deg}^-(x)}$
Preferential attachment ( $v$ to $w$ )	$\text{deg}^+(v) \cdot \text{deg}^-(w)$
Preferential Attachment ( $w$ to $v$ )	$\text{deg}^+(w) \cdot \text{deg}^-(v)$
Two-step paths ( $v$ to $w$ )	$ \text{paths}^2(v, w) $
Two-step paths ( $w$ to $v$ )	$ \text{paths}^2(w, v) $
Two-step paths ratio	$\frac{ \text{paths}^2(v, w) }{ \text{paths}^2(w, v) }$

### Comparison of prediction mechanisms

We observe higher accuracy for the REV task than SYM, as REV is “easier” than SYM since we know more information about the edge  $(v, w)$ .

Comparing REV- $v$  to REV- $w$ , we see REV- $w$  obtains higher accuracy, suggesting that when trying to predict the existence  $(w, v)$  of given  $(v, w)$ , knowing about properties of  $w$  is more

valuable than knowing properties of  $v$ .

Note that  $\text{SYM}^-$ ,  $\text{REV}^-$ ,  $\text{REV}-w^+$  and  $\text{REV}-v^-$  are such poor predictors that simply predicting that everything was reciprocated (or unreciprocated) would have been better.

### Comparison of methods of prediction

**Trends** On the whole, outdegree-indegree ratio and the two-step paths ratio are the best indicators of reciprocity. In fact, outdegree-indegree ratio alone already achieves accuracy to within  $\pm 5\%$  of a decision tree using every feature.

**Sending and receiving** When we look at features using one of the four mechanisms, we find that for the majority of the features larger values indicate reciprocity is more likely to occur. However, the *smaller* the outdegree-indegree ratio, the more likely reciprocation occurs. In other words, a large denominator and small numerator in  $\frac{\text{deg}^+(v)}{\text{deg}^-(v)} / \frac{\text{deg}^+(w)}{\text{deg}^-(w)} = \frac{\text{deg}^+(v) \text{deg}^-(w)}{\text{deg}^-(v) \text{deg}^+(w)}$  is a good indicator of reciprocation. A large denominator and small numerator indicate that  $v$  has many in-links and few out-links and that  $w$  has many out-links and few in-links. This suggests that  $v$  has higher “status” than  $w$  and hence increases the probability that  $w$  links to  $v$ .

Interestingly, separating the numerator and denominator from the outdegree-indegree ratio above, which corresponds to our two preferential attachment features, leads to very different results. While a small numerator does reasonably well (preferential attachment ( $v$  to  $w$ )), a large denominator does not (preferential attachment ( $w$  to  $v$ )) and performs only marginally better than chance. It is reasonable that the numerator and denominator would individually perform worse than the ratio, since the ratio takes into account both of them. However, the fact that the denominator performs so poorly is surprising because a large denominator suggests that  $v$  has a higher status than  $w$ ; this could increase the chance that  $w$  links to  $v$  even if  $w$  were to randomly link to others. On the other hand, a small numerator provides some information about status but not about increased reciprocation under random linking. The fact that a small numerator is more important than a large denominator suggests that status, as measured by the number of in- and out-links, is a powerful predictor of reciprocity.

If we consider only edges  $E_= = \{(v, w) \mid \text{deg}^-(v) = \text{deg}^-(w)\}$  (edges connecting nodes of equal degree), the accuracy of the outdegree ratio feature increases to 0.811. This result is comparable to what we get for the outdegree-indegree ratio feature.

Table 4.2: Indegree performance - different methods

<b>Mechanism</b>	<b>val<sub>OPT</sub> (Percentile)</b>	<b>Accuracy</b>
<i>Indegree ratio</i>		
SYM <sup>+</sup>	0.256 (40)	0.702
SYM <sup>-</sup>	-	-
REV <sup>+</sup>	0.414 (46)	0.759
REV <sup>-</sup>	-	-
<i>Indegree of <math>v</math> or <math>w</math></i>		
REV- $w$ <sup>+</sup>	-	-
REV- $w$ <sup>-</sup>	74 (61)	0.731
REV- $v$ <sup>+</sup>	61 (60)	0.582
REV- $v$ <sup>-</sup>	-	-

**REV- $v$  vs. REV- $w$**  REV- $w$  performs better than REV- $v$  on almost all features, and where REV- $v$  performs better, the difference is not as great. The fact that information about  $w$  is more useful than information about  $v$  suggests a contrast for various domains of potential application: for example, if we think of  $v$  as sending information to  $w$  via the  $(v, w)$  communication link (consider for example a marketer  $v$  contacting a potential customer  $w$ ), then we find that knowledge of the recipient ( $w$ ) tells us more about the probability of a response than knowledge of the sender ( $v$ ).

#### 4.4.2 Decision tree analysis

We can also combine subsets of features and evaluate their performance by randomly splitting the edges in  $E_{10}^r \cup E_{10}^u$  into two sets and performing 2-fold cross-validation. We use the ID3 algorithm to train the decision tree classifiers, and because the val features are continuous, we quantize each feature into deciles (dividing the data equally into tenths) to reduce computation time.

We consider the following combined sets of features, as well as each set individually:

1. **All** (sets 1-4) – every single feature was considered.
2. **All ratio** (sets 2,3,4) – all features that used ratios were considered.
3. **All absolute** (sets 1,3,4) – this allows us to see how using only “absolute” features affects accuracy.



Table 4.3: Reciprocity Prediction Feature Performance: Individual (REV)

Feature	val <sub>OPT</sub> (Percentile)	Accuracy
Indegree ratio	0.414 (46)	0.759
Outdegree ratio	0.667 (43)	0.628
Incoming message ratio	0.333 (48)	0.772
Outgoing message ratio	0.905 (46)	0.547
Incoming message-indegree ratio	0.650 (39)	0.569
Outgoing message-outdegree ratio	0.791 (33)	0.615*
Outdegree-indegree ratio	1.72 (53)	0.820*
Mutual neighbors (in)	10 (61)	0.552
Mutual neighbors (out)	8 (51)	0.580
Jaccard’s coefficient (in)	0.0345 (48)	0.684
Jaccard’s coefficient (out)	0.0637 (55)	0.660
Adamic/Adar	1.94 (55)	0.561
Two-step paths ( $v$ to $w$ )	6 (59)	0.517*
Two-step paths ( $w$ to $v$ )	5 (51)	0.657
Two-step paths ratio	0.556 (52)	0.760
Preferential attachment ( $v$ to $w$ )	10230 (58)	0.687*
Preferential attachment ( $w$ to $v$ )	2610 (37)	0.534*

Table 4.5 shows the accuracy of the trees and the most important attribute for the different sets of features. We find that using only degree/message features (set 1) performs as well as using all absolute features (sets 1, 3, 4). The two-step paths ratio alone obtains an accuracy of 0.760, while the decision tree for link prediction only manages 0.739. This can be attributed to inaccuracies introduced while quantizing the continuous features. Furthermore, features commonly used for link prediction yield a tree of lower accuracy than other features, providing evidence that the problem of reciprocity prediction is different from link prediction. Whenever the outdegree-indegree value or ratio was included in the feature vector, it was the single most important variable.

If we only consider  $E_{=}$ , node pairs with equal indegree ( $|E_{=}| = 16,311$ ), the accuracy of All ratio drops to 0.776. This suggests that predicting reciprocity becomes considerably more difficult as we lose the ability to differentiate between nodes of different status or indegree.

Table 4.4: Reciprocity Prediction Feature Performance: Individual (REV- $v$ ,REV- $w$ )

Feature	val <sub>OPT</sub> (Percentile)	Accuracy
Indegree ( $v$ )	61 (60)	0.582
Indegree ( $w$ )	148 (61)	0.731*
Outdegree ( $v$ )	25 (14)	0.506*
Outdegree ( $w$ )	105 (60)	0.647*
Incoming messages ( $v$ )	619 (53)	0.637
Incoming messages ( $w$ )	1802 (54)	0.733*
Outgoing messages ( $v$ )	906 (51)	0.542
Outgoing messages ( $w$ )	506 (17)	0.524*
Incoming message-indegree ( $v$ )	9.4 (41)	0.596
Incoming message-indegree ( $w$ )	9.12 (30)	0.535
Outgoing message-outdegree ( $v$ )	13.2 (50)	0.523
Outgoing message-outdegree ( $w$ )	8.14 (36)	0.661
Outdegree-indegree ( $v$ )	1.28 (53)	0.679*
Outdegree-indegree ( $w$ )	0.747 (50)	0.777

Table 4.5: Decision Tree Accuracy

Set	Accuracy	Top-level attribute
Degree/message (1)	0.832	Outdegree-indegree ( $w$ )
Degree/message ratio (2)	0.861	Outdegree-indegree ratio
Two-step hops (3)	0.796	Two-step paths ( $w$ to $v$ )
Link prediction (4)	0.739	Two-step paths ratio (directed)
<i>Combined</i>		
All ratio (2,3,4)	0.861	Outdegree-indegree ratio
All absolute (1,3,4)	0.832	Outdegree-indegree ( $w$ )
All (1-4)	0.862	Outdegree-indegree ratio

Again, as the indegrees of  $v$  and  $w$  were equal in every pair  $(v, w)$ , it is not surprising that the outdegree ratio is the most important feature.

Table 4.6: Logistic regression – relative degree/message-based features

Feature	$\beta$	p value
Indegree ratio	0.0101903	$< 2 \times 10^{-16}$
Outdegree ratio	0.0005775	<del>0.2545</del>
Incoming messages ratio	0.0230161	$< 2 \times 10^{-16}$
Outgoing messages ratio	-0.0047152	$< 2 \times 10^{-16}$
Incoming messages-indegree ratio	-0.0005545	<del>0.0798</del>
Outgoing messages-outdegree ratio	-0.0049387	$< 2 \times 10^{-16}$
Outdegree-indegree ratio	<b>-0.0562983</b>	$< 2 \times 10^{-16}$

## Performance

Classifying features based on their computation time, the two-step hop and link prediction features (excluding preferential attachment) take more than two orders of magnitude longer to compute than all other features (500 times as long for  $k = 10$ ). If we only use the other features in prediction, we still obtain an accuracy of 0.862, similar to what we obtained above when we used all features. Therefore, it appears sufficient to use only these features in practical applications.

## Effect of $k$ on accuracy

As  $k$  increases, the proportion of edges that are defined as reciprocated increases, and naturally accuracy also increases (0.8818 for  $k = 20$  and 0.9032 for  $k = 50$ ). If we instead take equal numbers of reciprocated and unreciprocated edges, giving a baseline accuracy of 0.500, accuracy gradually increases from 0.836 for  $k = 10$  to 0.846 for  $k = 30$ .

### 4.4.3 Regression analysis

We also used a logistic regression model on subsets of features, where  $f(z) = \frac{e^z}{e^z + 1}$ ,  $z = \beta_0 + \beta F$ ,  $f(z)$  is binary (1 when an edge is reciprocated, 0 otherwise) and  $F$  is the vector of features. The results are shown in tables 4.6—4.9, where struck-out p-values indicate insignificant features. Here, the two-step paths ( $v$  to  $w$ ), two-step paths ratio, and Jaccard (out) features are most significant when simultaneously using all features for classification.

Table 4.7: Logistic regression – two-step hop features

Feature	$\beta$	p value
Mutual neighbors (in)	-0.0117269	$< 2 \times 10^{-16}$
Mutual neighbors (out)	0.0180579	$< 2 \times 10^{-16}$
Two-step paths ( $v$ to $w$ )	-0.1193624	$< 2 \times 10^{-16}$
Two-step paths ( $w$ to $v$ )	0.1296081	$< 2 \times 10^{-16}$

Table 4.8: Logistic regression – All ratio

Feature	$\beta$	p value
Indegree ratio	0.0120256	$< 2 \times 10^{-16}$
Outdegree ratio	-0.0015554	0.005739
Incoming messages ratio	0.0145437	$< 2 \times 10^{-16}$
Outgoing messages ratio	-0.0043189	$< 2 \times 10^{-16}$
Incoming messages-indegree ratio	0.0048525	$< 2 \times 10^{-16}$
Outgoing messages-outdegree ratio	-0.0046674	$< 2 \times 10^{-16}$
Outdegree-indegree ratio	-0.0301592	$< 2 \times 10^{-16}$
Mutual Neighbors (in)	-0.0279290	$< 2 \times 10^{-16}$
Mutual Neighbors (out)	0.0147103	$< 2 \times 10^{-16}$
Two-step paths ( $v$ to $w$ )	<b>-0.0530463</b>	$< 2 \times 10^{-16}$
Two-step paths ( $w$ to $v$ )	0.0182572	$< 2 \times 10^{-16}$
Two-step paths ratio	<b>0.0394657</b>	$< 2 \times 10^{-16}$
Jaccard (in)	-0.0238541	$< 2 \times 10^{-16}$
Jaccard (out)	<b>0.0572358</b>	$< 2 \times 10^{-16}$
Adamic-Adar	-0.0001424	<del>0.881637</del>
Preferential attachment ( $v$ to $w$ )	0.0010837	0.000627
Preferential attachment ( $w$ to $v$ )	-	-

## 4.5 Twitter as a superposition of networks

### 4.5.1 (Un)reciprocated subgraph analysis

We also analyze how various properties of the subgraphs  $G_n$ , as well as the edge sets  $E_k^r$  and  $E_k^u$ , vary as we adjust  $n$  and  $k$ .

**Reciprocated and unreciprocated edges** We observe that the frequency of reciprocated edges is approximately 2 to 3 times that of unreciprocated edges, and the proportion of reciprocated edges increases as  $n$  and  $k$  increases (Fig. 4.1, 4.2). While reciprocated communication is the dominant form of interaction, we also see a significant number of unreciprocated interactions, indicating that a significant number of relationships on Twitter are unbalanced. This could occur when a user of lower status repeatedly invokes the name of a more influential user (of higher status) through @-mentioning, as suggested in the introduction.

**Reciprocated and unreciprocated nodes** A significant proportion of nodes take part in both reciprocated and unreciprocated interactions, and while a majority of nodes have reciprocated interactions, only a small proportion have purely unreciprocated interactions. This indicates that while there are two distinct types of relationships occurring on Twitter, they do not correspond to two distinct types of users. The fact that it is rare to find active Twitter users taking part in only unreciprocated interactions suggests that social (reciprocal) relationships are associated with an active and continued use of the site.

We can also see this in a scatter plot of the number of users  $v$  with each of three types of interaction (Fig. 4.5): (i) reciprocated interactions with both  $(v, w)$  and  $(w, v)$  present; (ii) unreciprocated “out-going” interactions with only  $(v, w)$  present; and (iii) unreciprocated “in-coming” interactions with only  $(w, v)$  present. We thus differentiate between both ends in an unreciprocated edge ( $v \xrightarrow{k} w$  and  $w \xrightarrow{=0} v$ ), where a user could play the role of  $v$  if she’s not replied to, or the role of  $w$  if she doesn’t reply. From the plot, we can see that the types with the greatest numbers of associated nodes are those with only reciprocated interactions.

**Clustering coefficient remains relatively stable as  $n$  and  $k$  vary** The clustering coefficient is much larger in the subgraph of reciprocated edges, which corresponds to the natural notion that reciprocated edges represent more social activity with a larger density of triangles. The fact that these quantities are stable as we change  $n$  and  $k$  suggests that the network properties of these subgraphs do not change significantly even if we sample from a relatively smaller population of all users (Fig. 4.3).

**Connected component remains stable as  $n$  varies, but decreases as  $k$  increases** The graphs corresponding to  $E_k^r$  and  $E_k^u$  have giant components for relatively low values of  $k$ . However, the

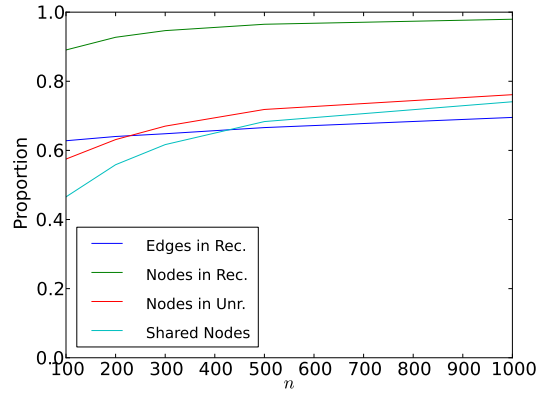


Figure 4.1: Proportion of nodes or edges (varying  $n$ )

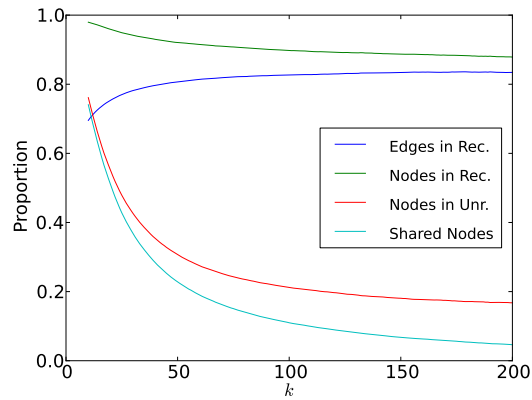


Figure 4.2: Proportion of nodes or edges (varying  $k$ )

size of the largest component shrinks rapidly once  $k$  passes a particular range (roughly between 50 and 100). (Fig. 4.4). This hints at a kind of qualitative transition in the structure of the network as a function of message volume, with the edges representing more than 100 communications each unable to sustain a very large component on their own.

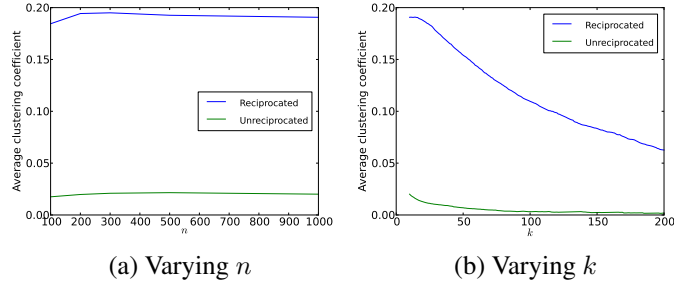


Figure 4.3: Clustering coefficient

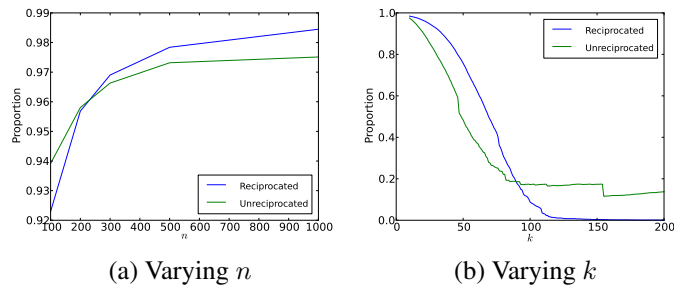


Figure 4.4: Proportion in largest connected component

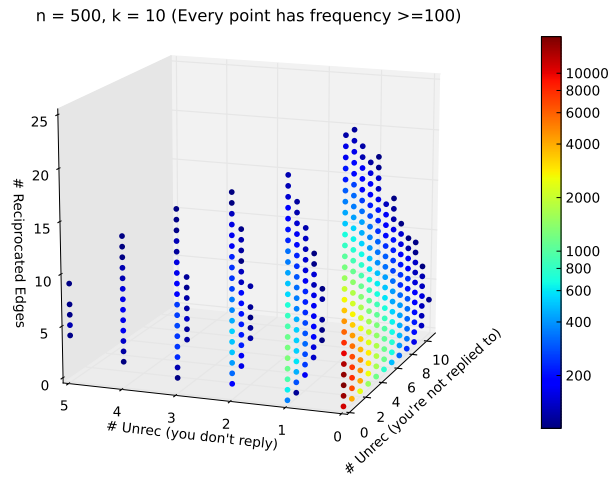


Figure 4.5: Scatter plot of users' interaction types

Table 4.9: Logistic regression - All

Feature	$\beta$	p value
Indegree ratio	0.0041791	$6.09 \times 10^{-8}$
Outdegree ratio	0.0046914	$1.28 \times 10^{-13}$
Incoming messages ratio	0.0029794	0.000125
Outgoing messages ratio	0.0033361	$3.10 \times 10^{-10}$
Incoming messages-indegree ratio	0.0040884	$1.11 \times 10^{-14}$
Outgoing messages-outdegree ratio	-0.0015075	0.006283
Outdegree-indegree ratio	-0.0057958	$< 2 \times 10^{-16}$
Indegree (v)	0.0050520	$4.30 \times 10^{-11}$
Indegree (w)	-0.0089197	$< 2 \times 10^{-16}$
Outdegree (v)	-0.0063247	$< 2 \times 10^{-16}$
Outdegree (w)	0.0035881	$3.58 \times 10^{-7}$
Incoming messages (v)	0.0063390	$< 2 \times 10^{-16}$
Incoming messages (w)	-0.0179975	$< 2 \times 10^{-16}$
Outgoing messages (v)	-0.0070869	$< 2 \times 10^{-16}$
Outgoing messages (w)	0.0095572	$< 2 \times 10^{-16}$
Incoming message-indegree (v)	-0.0023250	$1.11 \times 10^{-5}$
Incoming message-indegree (w)	-0.0004044	<del>0.42781</del>
Outgoing message-outdegree (v)	0.0007430	<del>0.175454</del>
Outgoing message-outdegree (w)	0.0024155	$2.54 \times 10^{-5}$
Outdegree-indegree (v)	-0.0110324	$< 2 \times 10^{-16}$
Outdegree-indegree (w)	0.0218874	$< 2 \times 10^{-16}$
Mutual Neighbors (in)	-0.0194635	$< 2 \times 10^{-16}$
Mutual Neighbors (out)	0.0050245	$< 2 \times 10^{-16}$
Two-step paths (v to w)	-0.0462950	$< 2 \times 10^{-16}$
Two-step paths (w to v)	0.0167156	$< 2 \times 10^{-16}$
Two-step paths ratio	0.0440107	$< 2 \times 10^{-16}$
Jaccard (in)	-0.0398243	$< 2 \times 10^{-16}$
Jaccard (out)	0.0561815	$< 2 \times 10^{-16}$
Adamic-Adar	0.0111504	$< 2 \times 10^{-16}$
Preferential attachment (v to w)	0.0009537	0.003002
Preferential attachment (w to v)	-	-



## Chapter 5

# Duolingo: Massive Online Language Education

Duolingo is the world's largest language learning platform built on the premise that high-quality language education should be available to everyone. As of November 2014, over fifty-five million people have started learning a foreign language with Duolingo. One unique aspect of Duolingo is that students (we use the term user and student interchangeably to describe someone who uses Duolingo) can also learn by translating actual content from the internet. This not only exposes students to more complex sentences that are representative of what they will encounter when using the language, but also enables Duolingo to produce high-quality translations of documents. We provide a brief overview of the core components of the Duolingo platform, with an emphasis on the social features of the system. A more comprehensive overview of Duolingo, with a particular emphasis on the translation component, can be found in Severin Hacker's PhD thesis [30].

### 5.1 Platform Overview

The Duolingo platform consists of many components that come together in a production system that serves more than 1.1 million students every day. It is available at [www.duolingo.com](http://www.duolingo.com) as well as on three mobile platforms (iOS, Android, and Windows Phone) and is used in 134 countries around the world (counting countries with more than 100 daily active users as of November 2014).

### 5.1.1 Teaching Objectives

Duolingo aims to teach a core vocabulary that is approximately the 3,000 most common words in a language. This enables students who complete a course in Duolingo to have a reasonable understanding of most text they encounter in everyday situations. For example, a student who completes the English course on Duolingo can expect to know more than 90 percent of text they encounter in everyday writing.

This vocabulary is split into units call *skills* that are arranged in a *skill tree*. Skills introduce new vocabulary using sentences constructed from words that the student has already learned, forcing a natural progression through the skill tree. Sentences for words taught in each skill build off of each other: for example, the first skills teach basic nouns such as {*man, woman, boy, bread, water*}, subjects such as {*I, You, She*}, and conjugations of infinitives such as *to be, to eat, to drink*. Very quickly nontrivial vocabulary is introduced and students start encountering more complicated sentences. Previously, the skill tree was closer to an actual tree with node dependencies in which a skill would unlock only if all of its dependent skills were also unlocked. A student must now complete all of the skills in a given level before accessing *any* of the skills in the next row. Figure 5.1 illustrates the structure of the skill tree in German.

### 5.1.2 Student Learning History

Duolingo tracks the vocabulary learned by each student, as well as statistics such as how many times a student has seen each word, gotten each word correct, and when they last saw each word. These statistics are used in a model that we trained to estimate how likely a student will get a word correct when presented a challenge containing that word. For each word in a student's vocabulary we estimate a probability that the student will get the word correct in a challenge. We call this probability the *strength* of the word, use it to indicate how well a student remembers the collective vocabulary in each skill. This *strength* model has been trained on many tens of millions of observations of student performance in the lessons. Skills are assigned a strength that is the average strength of words taught in the skill. These skill strengths correspond to the strength bars seen around each skill. If the skill strength is above a certain threshold the skill turns golden. This mechanism is powerful as students strive to keep their skill tree as golden as possible. Several times a bug has been introduced that prevents students from fully strengthening

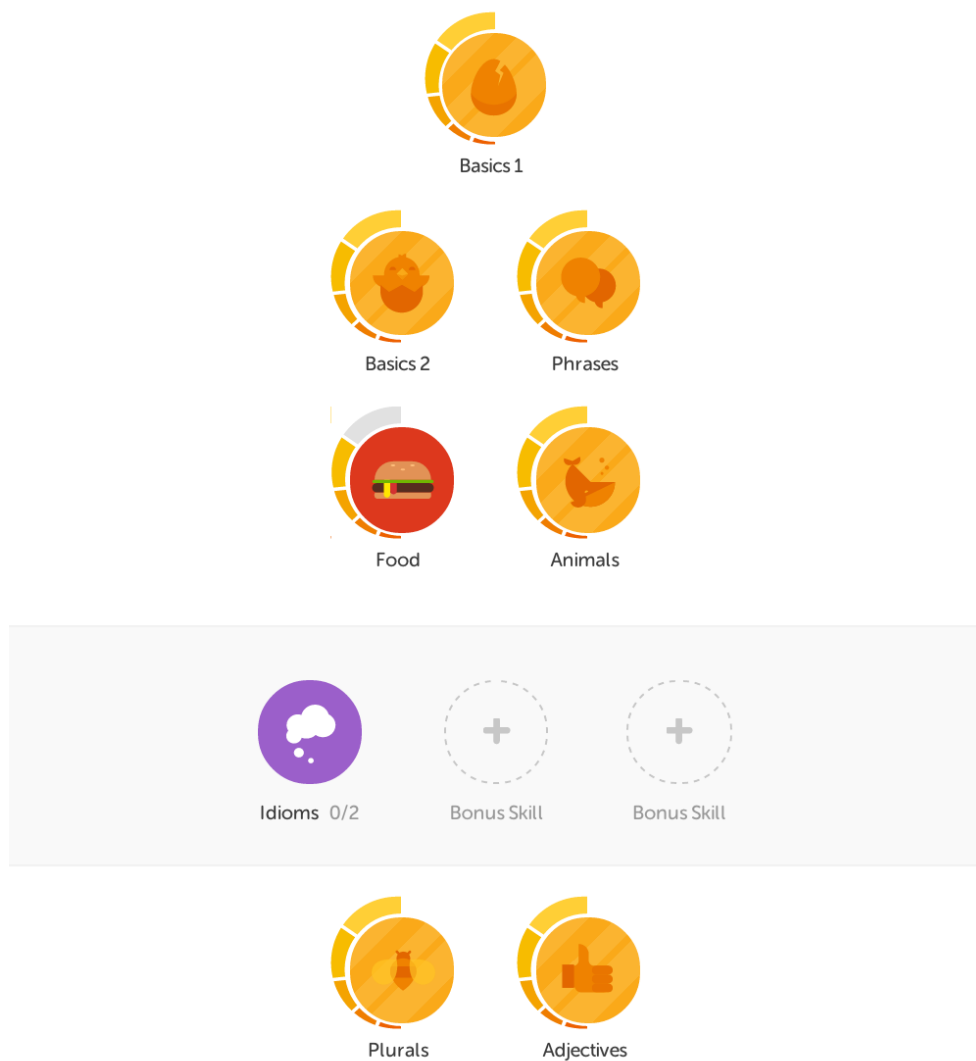


Figure 5.1: The start of the German skill tree.

their skills and this problem is quickly reported to us.

### 5.1.3 Personalized Lessons

A distinctive characteristic of Duolingo is that each student receives personalized lessons and practice that are dynamically generated based on the student's learning history. The core algorithm, which we call *Session Generator*, combines a student's learning history with the corpus

## You strengthened Qualifiers, Frequency and Adverbs 1



Keep those **strength bars** full as words fade from your memory.

Figure 5.2: Strengthening skills in practice.

of language data created in the Incubator to output lessons and practice. Generating lessons is a far more constrained problem because a particular set of words will be taught regardless of the strength of the student's vocabulary. Practice sessions are generated by picking words that have the greatest expected increase in strength. Because of how the current strength model works, this effectively means that Session Generator picks the words a student knows least, but this may not always be the case should the model change. As session generator is the most integral part of Duolingo, modifications to Session Generator are heavily scrutinized in code review and are introduced through A/B testing.

### 5.1.4 Intelligent Feedback

Student responses are algorithmically examined for mistakes. We can recognize mistakes such as incorrect verb conjugation, incorrect gender, and wrong word order. Errors such as typos are handled in a more lenient manner where we normally will not penalize a user for common mistakes such as 'teh'. The output of this grading algorithm is used to communicate what mistakes the student made and to update a student's knowledge model appropriately. We do not simply mark all words in a sentence as correct or incorrect based on the response's overall correctness (early in Duolingo this was actually done, however). This method is not yet ideal as certain mistakes, such as using the incorrect gender of a noun in German, penalize the gendered article and not the noun itself.

## 5.2 Game Mechanics

Duolingo uses a variety of techniques borrowed from games to make the learning experience engaging and fun:

### 5.2.1 eXperience Points and Levels

Students gain experience points (XP) for completing lessons, practicing, or contributing to real-world translations. These XP are used to rank users in the leaderboard. The leaderboard features a weekly, monthly, and all-time category so that everyone, especially new users, are able to compete from the same starting point each week or month. Veteran users can pride themselves on having an extremely high cumulative XP.

Additionally, the XP are also used to determine one's *level* in each course they are learning. There are currently twenty-five levels structured such that additional levels can be added later. Completing a course does not consist of enough work to reach the highest level. This is to encourage students to continue refreshing their skills and contribute to the Immersions. Furthermore, XP for the sake of level calculation is course-specific: students learning multiple courses will likely have different levels in each course.

### 5.2.2 The Coach and Streaks

Learning a language takes consistent practice over long periods of time. Duolingo has two features that emphasize the importance of regular practice. In early 2013 we introduced *streaks*, which measure the the number of consecutive days a student has been active on Duolingo. Email reminders emphasize the student's current streak and use phrases such as "You're on fire!" for encouragement. Simply getting one experience point per day is sufficient to extend one's streak. While this does get students to return to Duolingo, it is not necessarily good for the learning experience. The coach takes the streak concept one step further by requiring students to set a daily XP goal. Different levels from 'Casual' (10 XP/day) to 'Insane' (50 XP/day) provide students a range for how dedicated they want to be. In general, it takes about four minutes (the average time to complete a lesson) to earn 10 XP.

Streaks are very popular with the community; an entire discussion thread has been started to bring attention to students who have streaks over one year. Each student features *flair* consisting

of their course levels and streak as a badge of honor.

AlexisLinguist 🇪🇸 25 🇺🇸 13 🇬🇧 13 🇮🇹 12 🇫🇷 7 🇸🇪 5 🇩🇰 5 🔥 217

Figure 5.3: Flair for a student in the discussion forums.

The coach has superseded the streak as the preferred way to motivate students and is now available on all platforms.

### 5.2.3 Lingots

*Lingots* are Duolingo's virtual currency that can be used to buy items from the *Lingot store*. Completing skills and maintaining streaks are two ways in which students can earn Lingots. Every day more than one million Lingots are awarded to students. Items that can be purchased include outfits for the Coach mascot, bonus skills, and additional features such as a timed practice mode or progress quiz. On a given day approximately 3% of students will purchase an item from the Lingot store.

## 5.3 Social Features in Duolingo

Duolingo has several social feature that allow students to interact with each other. The first feature is the ability to *follow* other learners, similar to how one follows other users on Twitter. When a student A follows student B, A can see B's progress in their activity stream. Some activities that appear in the activity stream include commenting in the discussion forums, reaching a new level, or learning a new skill. Students can comment on activities or *like* them, and these actions are visible to all Duolingo students. The activity stream has not been integrated into the mobile applications. These features are shown in Figure 5.4. Nearly 10% of students on the website click the stream each day.

In addition to appearing in the activity stream, the overall learning progress for followed students also appear in a *leaderboard*. The leaderboard shows one's weekly, monthly, and all-time progress compared to those users they follow. This leaderboard view is present whenever a user visits their home screen on the website; in the mobile applications users must navigate to

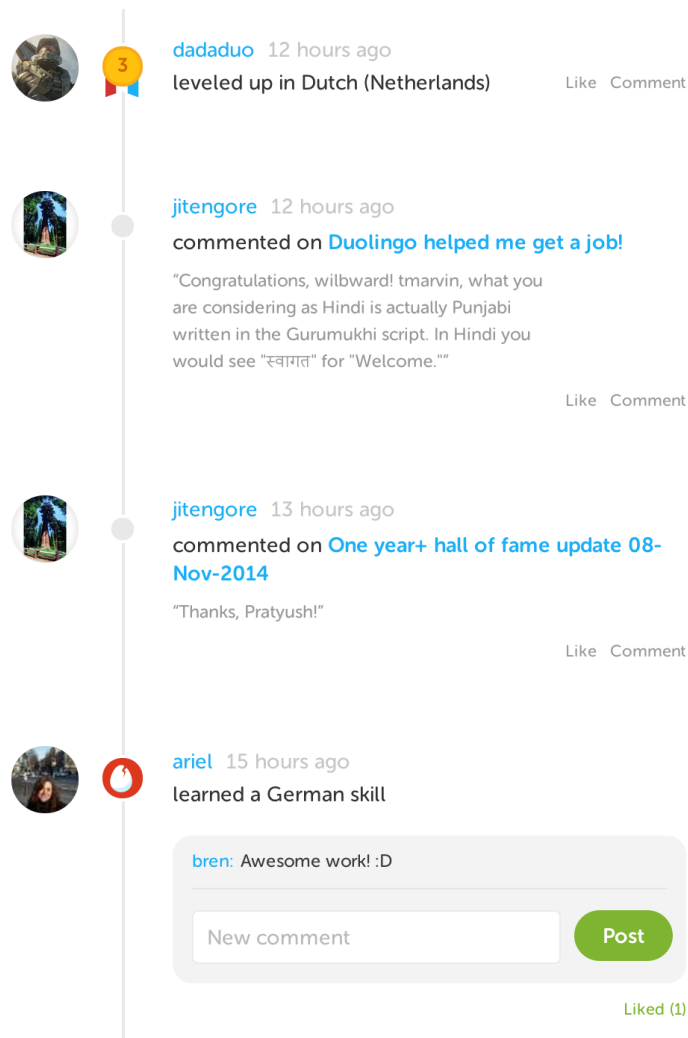


Figure 5.4: The activity stream in Duolingo.

their profile to view the leaderboard. Figure 5.5 illustrates how the leaderboard is integrated into the website.

When someone is followed by another student we notify them in two ways. First, we send an email to the followed user telling them they are now followed and some information about the user following them. Additionally, the next time the followed user goes to their home screen on the website, a *notification area* will tell them that they are now being followed. This link can be easily reciprocated via a link in the email or a button in the notification area. The follow-back behavior of users will be discussed in Chapter 6.

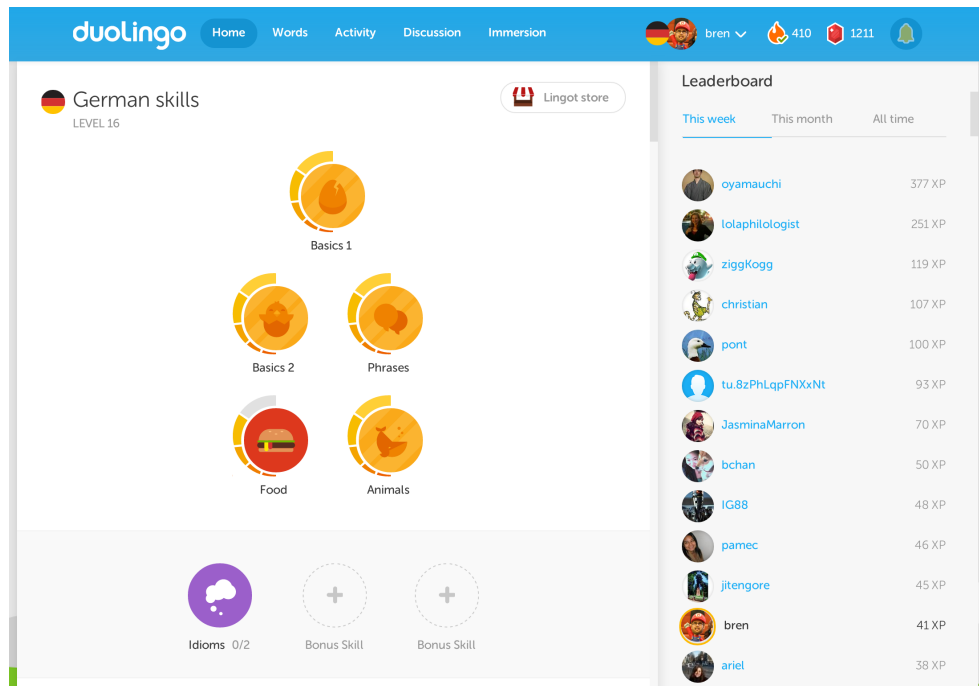


Figure 5.5: The Duolingo leaderboard.

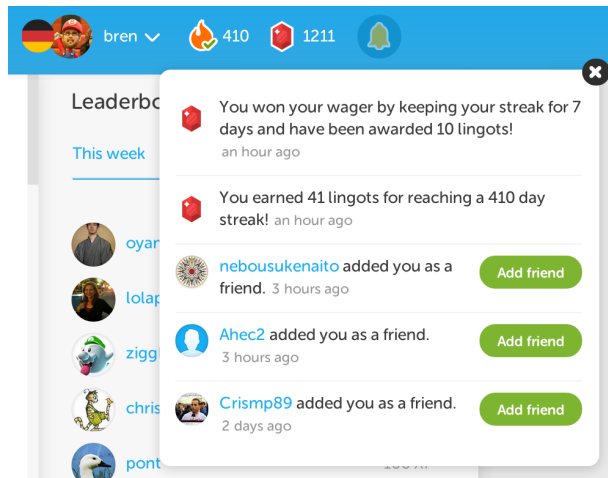


Figure 5.6: The notification area displays recent events.

### 5.3.1 Finding Other Students on Duolingo

Students can find their friends on Duolingo in several ways. The easiest is to associate a Facebook account with a Duolingo account. Doing so lets a student see which of their friends are already using Facebook and send invites to friends who are not yet using Duolingo.

Using this feature several thousand student invites are sent each day. Students can also search



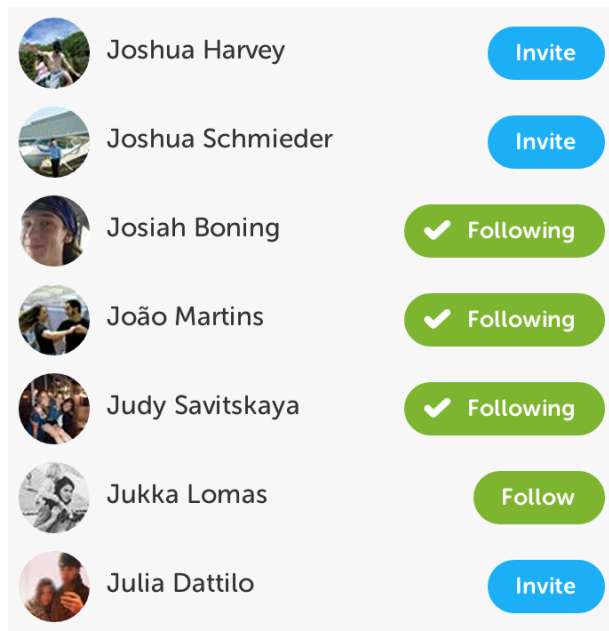


Figure 5.7: Using Facebook to find friends on Duolingo.

for each other by name or email address. Searching by name is done in a naive manner and is not particularly effective. Improving this functionality by suggesting students to follow based on the social graph, or by providing better search tools, is one way we could increase user following in the future.

## 5.4 Community Features in Duolingo

Duolingo not only allows users to interact directly with each other via social features, but also in a larger scope via its *forums* and online presence at Facebook and Twitter. The Duolingo forums serve as a way for users to ask questions about languages and the product. Approximately 8% of students using the website visit the forums each day; the forums are not available in the mobile applications. A total of 90,895 posts have been started by 49,864 students. Students can subscribe to different forum topics and are automatically subscribed to the ‘Duolingo’ topic in their native language and any course they add.

The forums allow students to ask language questions, share their learning experiences, or report technical issues. Within a comment thread students can reply to each other in a standard nested manner. For both design and technical simplicity, we bound the nested reply depth to five

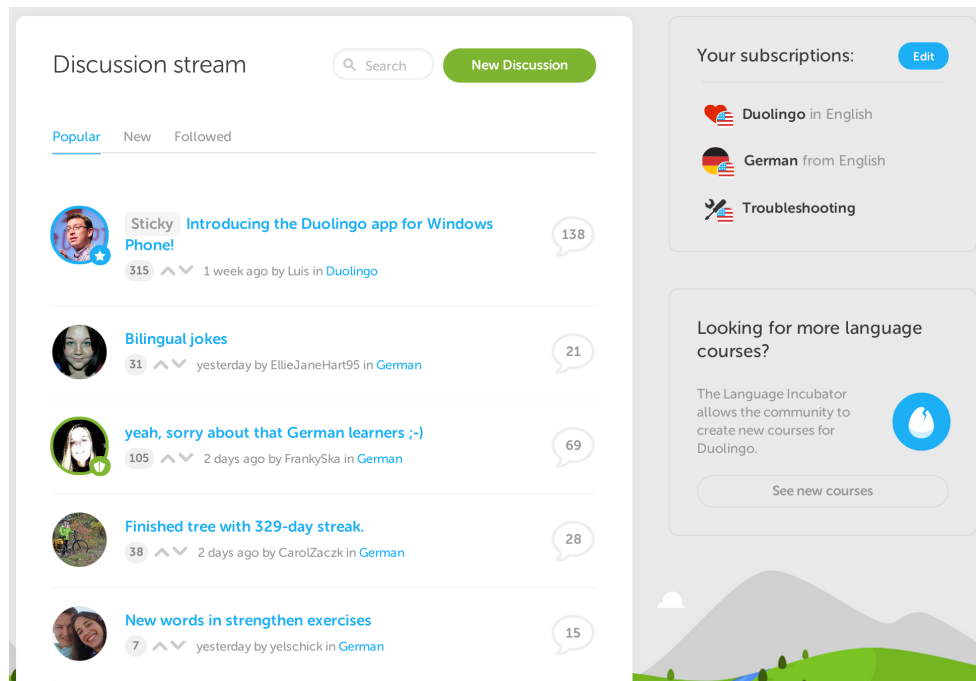


Figure 5.8: The main discussion view shows posts from subscribed topics.





comments. Students can vote comments up or down, and these votes are used to order the replies within the thread. Additionally, students can give their Lingots to comment authors as a token of appreciation. Reddit, a ‘platform for online communities’ has a similar set of features within their comment system.

## 5.5 Measurement and Experimentation




Developers at Duolingo measure almost every student action through several methods. First are the access logs that keep track of which requests students make. These are most useful for diagnostics purposes and can be used to make interesting graphics of how our students are distributed around the world. The access logs are not sufficient for understanding user behavior as some actions are taken without making any requests to the backend services. Moreover, they are not in an ideal format for analysis and aggregation. We use an online service called Mixpanel [61] to track student *actions* in the apps and on the website. Rather than record page views, specific actions are captured that give a detailed view of user behavior. For example, we record activity such as when the application is loaded, a skill is visited, a lesson is started, and whether a lesson is successfully or unsuccessfully completed.

## 21 Comments





---

 **Maraplu** 🇧🇪 15 🇩🇰 12 🇩🇪 10 🇮🇹 9 🇸🇪 6 🔥 87  
I only know a really bad one.  
According to Freud, what comes between fear and sex?  
Fünf  
79   Reply Edit Delete Give Lingot  20 • yesterday




---

 **Gevgever** 🇩🇪 16 🇧🇪 11 🇪🇸 6 🔥 1  
This is amazing. It's like the first time you hear "seven ate nine"  
10   Reply Edit Delete Give Lingot • yesterday

---

 **\_Serpico\_** 🇩🇪 7 🇸🇪 5 🇮🇹 4 🔥 9  
Have all my lingots~! /o/  
3   Reply Edit Delete Give Lingot  10 • yesterday

---

 **SOROUGH** 🇩🇪 13 🇺🇸 5 🔥 11  
Oh my god this is gold.  
Even Freud's name sounds like "pleasure".... ;)  
1   Reply Edit Delete Give Lingot • 17 hours ago

---




 **writingfish** 🇩🇪 9 🇮🇹 4 🇺🇸 3 🔥 10  
haha, really. At first I think that will be the point of joke  
0   Reply Edit Delete Give Lingot • 15 hours ago

Figure 5.9: A comment thread showing flair and additional social features.

By using these activities several key metrics are measured and monitored over time. The overarching metric we monitor and optimize is student retention; students must actively engage in the service over time to successfully learn with Duolingo. Product changes are introduced via A/B test experiments in which we track the retention metrics and any other key metrics for the feature. At any given time we administer over twenty A/B tests and have developed a sophisticated framework for defining, implementing, and analyzing the outcome of an A/B

test. Mixpanel makes it easy to view such trends over time, partition students into different populations, and evaluate the outcome of A/B tests.

Finally, we produce over one-hundred gigabytes per day of backend logs that detail many different aspects of the system. For example, we track every update to student vocabularies. These logs are then used in a training algorithm for the vocabulary strength models. We also collect statistics about how frequently students get certain challenges incorrect to identify sentences that might have a data problem. Recently, we have changed Session Generator to use these challenge-level statistics to select which sentences to use.

## 5.6 Infrastructure

A common infrastructure running on Amazon Web Services [6] serves the learning application on all four platforms. It is written primarily in the Python programming language and responds to more than 400 million API requests every day. Approximately 200 modestly powered virtual machines (8 vCPUs, 15GB RAM) to run Duolingo. The architecture follows a modern multi-tier partitioning consisting of web-facing load balancing instances that proxy traffic to different services on the backend (application tier). Most of the applications in the Duolingo backend run independently; for example, the service that provides dictionary hints to students has very little overlap with the main application that powers the API. We run all of our applications in multiple *availability zones*, which is Amazon's terminology for a physically separated set of computing resources (data centers). In the event that one zone fails the others should have enough excess capacity to continue serving requests until corrective actions (such as launching more machines in the healthy zones) complete. Using this approach we have achieved a service uptime of 99.96% in 2014. A highly simplified architecture diagram is in figure 5.10.

Duolingo is still mostly a monolithic Python application where core components (such as generating lessons) have been factored out into self-contained services. While there is much debate over the strengths and trade-offs of a Service Oriented Architecture (SOA) system, we have found that a certain level of specialization with services allowed us to significantly improve performance. We discovered that the parts of the system which heavily use the relatively small (several gigabytes) collection of language data saw the biggest improvement by keeping this data in-memory instead of in a centralized data store (MySQL or Redis).

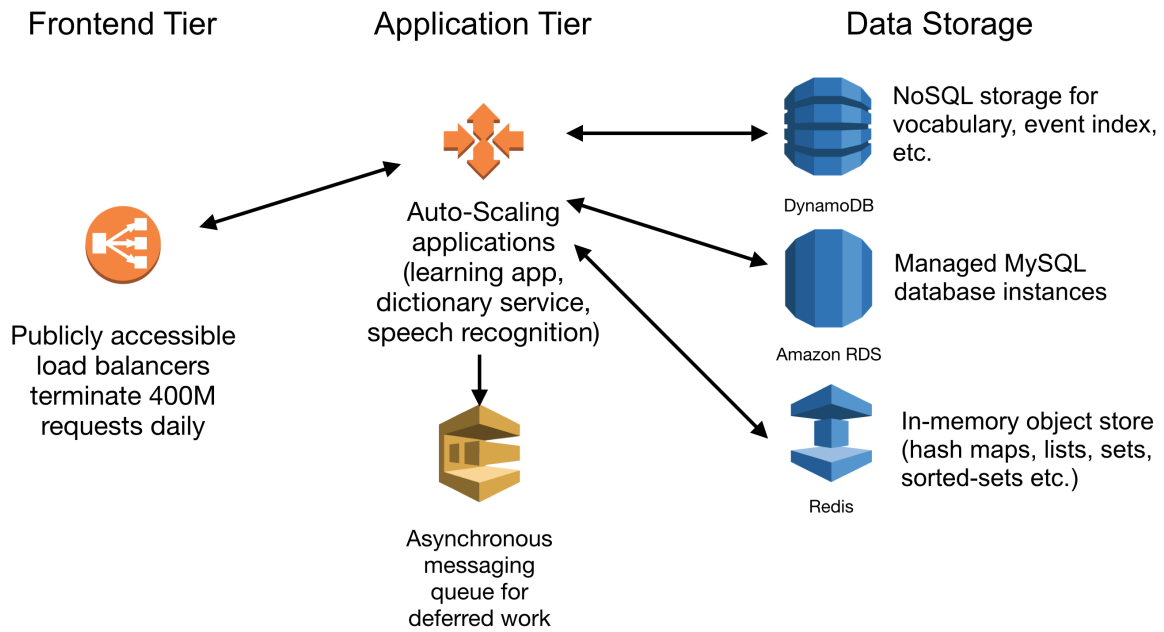


Figure 5.10: Simplified architecture diagram of Duolingo.

For data storage we primarily use two MySQL instances, Amazon’s DynamoDB NoSQL database, and Redis. When we first launched Duolingo we only used MySQL for storing ground truth and Redis for storing cached data. In December 2012 the approach of storing all data in MySQL was causing excessive load on the database. Specifically, the student vocabulary data wasn’t well suited to being stored in a single MySQL instance. Rather than sharding this data across multiple MySQL instances, we migrated the user vocabulary data to DynamoDB. DynamoDB provides very efficient data access for Duolingo vocabulary data (for a given user and course, get all vocabulary items), and more importantly, Amazon deals with scaling the underlying storage and compute resources as our needs increase. After the initial migration of 100 million vocabulary items, we have grown the amount of vocabulary data stored to nearly 13 billion items without any issues. After seeing the success of using DynamoDB for our large and frequently changing data sources we have migrated many other systems to DynamoDB with similar success.



# Chapter 6

## Duolingo's Social Network

This chapter explores the Duolingo student population across different dimensions as well as the structure and evolution of the social network over time.

### 6.1 Breaking down the Student Body

We consider partitioning the student population across the courses learned and platforms used. Over time these populations change as new courses are released; however, we will see that some trends remain unchanged over time. For much of Duolingo's history the first step was to create a new user account. Early in June 2014 we introduced a *delayed signup* capability so that students can try the application without registering with their email address or social network account. We call these users *trial students* and reserve a user account in the database for them so that their progress is maintained should they eventually convert to a registered user. There are many trial users and we are careful to take them into account when performing later analyses. Table 6.1 lists some statistics about the Duolingo student population.

Number of accounts	58,314,540
Number of registered users	46,416,394
Number of trial students	11,898,146
Facebook connected users	11,048,536

Table 6.1: Account statistics as of November 23, 2014.

The majority of students (56% between September 1 and October 30, 2014) accessed Duolingo on mobile devices through the mobile applications. This is important to note because some social

and community features are either absent or are not as prominently displayed in the mobile apps due to the limited screen sizes.

### 6.1.1 Languages learned

Among the 39 many courses offered as of November 26, 2014, more than half (22) teach English. Based on the courses we offer, as well as the demand for learning English, it is not surprising that English is the most popular language learned on Duolingo. Only 42% of students are native English speakers, leaving 8% of students studying a course in which neither the learning nor the native language is English. Table 6.2 lists the most popular languages taught on Duolingo.

English	49.7%
Spanish	19.6%
French	14.0%
German	8.5%
Italian	4.7%
Portugese	3.0%
All others	< 1% individually, <3% aggregated

Table 6.2: The most popular languages learned on Duolingo between September 1, 2014 and October 30, 2014

Among students learning English, the most common native languages are Spanish (41.1%), French (24.5%), German (16.0%), and Italian (10.28%). A long tail of native languages such as Dutch, Danish, Russian, and Japanese result from the many incubated courses offered. These statistics are not specific to the time range; the ordering and relative proportions of the most learned languages have remained the same for nearly a year.

Although students can learn multiple languages, we find that most students ‘enroll’ in a single course. Among all students in our system, 83.7% learn only one language (course). Those students with English as their native language have the highest fraction (23%) learning more than one language, with 3% learning 4 or more languages.

## 6.2 Network Structure

Like many social networks, Duolingo’s social network consists of a single large connected component, long-tailed in- and out-degree distributions, and small diameter. We consider the directed



relationship  $A$  follows  $B$  to be the directed edge  $A \rightarrow B$  in the social graph. Thus, the in-degree of a user  $B$  is the number of followers she has.

A summary of high-level statistics is found in Table 6.3.

Vertices	8,311,514
Edges	22,661,678
Number of users following someone	4,723,033
Number of users followed	7,447,116
Edges reciprocated	7,743,306
Edges where both users are connected to Facebook	13,897,279
Edges where following user is connected to Facebook	15,752,774
Edges where followed user is connected to Facebook	15,920,606
Number of users who are connected to FB following someone	2,305,072
Number of users followed who are connect to FB	4,636,434

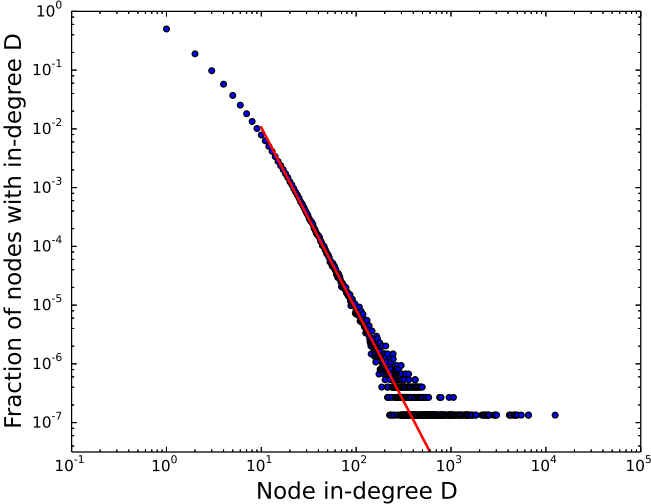
Table 6.3: Network statistics for the Duolingo network as of November 23, 2014.

There are a few interesting observations about these high-level statistics. Only 18% of users are present in the network, but this alone does not indicate that the social features are unpopular among students. Approximately 30% of users access Duolingo on the website, the platform that most prominently exposes social features. Only 20% of students that associate Facebook with their Duolingo account follow someone, yet nearly 49% of students following someone have connected Facebook to their accounts. Approximately 15% of edges in the network are reciprocated, which is close to the rate with which edges in Twitter are reciprocated. Based on the number of edges where one or both of the endpoints is connected to Facebook, it appears as though integration with Facebook is important to creating new edges. We show that this intuition is correct later in Section 6.3.

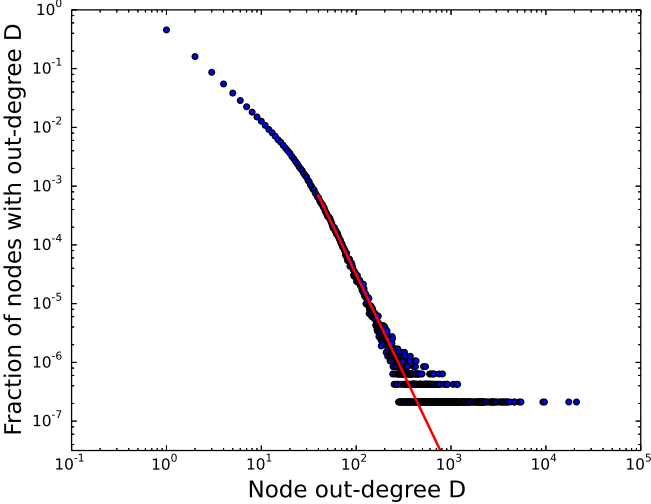
### 6.2.1 Degree distribution, connected components

The Duolingo social graph follows many of the usual patterns observed in other online networks. There is a long-tailed power-law for both the in- and out-degree distributions, a single large weakly connected component (LWCC) containing most nodes, and many small components that are not part of the LWCC. The in- and out-degree distributions are fitted using a maximum likelihood estimation library *powerlaw* written in Python [5, 69]. In figure 6.1 we show the degree distributions along with the fitted model. For the in-degree distribution we fit the tail where the

degree is at least 10 and find the exponent to be  $\alpha = -3.105$ . The out-degree distribution is fit for nodes with degree at least 40 and find the exponent to be  $\alpha = -3.375$ .



(a) In-degree (# followers) distribution.



(b) Out-degree (# following) distribution.

Figure 6.1: Degree distributions in Duolingo’s social network.

We note that a power-law fit on the out-degree distribution is not good if we fit for degree  $D = 10$  or greater. Using the powerlaw package we can compare multiple distributions and look at fitting the out-degree distribution to a power-law and log-normal distribution with minimum degree  $D_{min} = 10$ . A log-normal distribution with parameters  $\mu = 0.0513, \sigma = 1.44367$  provides a significantly better fit in this range, with the log-likelihood ratio exceeding 4000. In this

range ( $D_{min} = 10$ ) the best-fit power-law exponent is  $\alpha = -2.61112$ . Fitting only the tail of  $D \geq 40$  with both distributions shows that the power-law fit is actually better fit. Figure 6.2 compares the density function of the fitted distributions to the actual density.

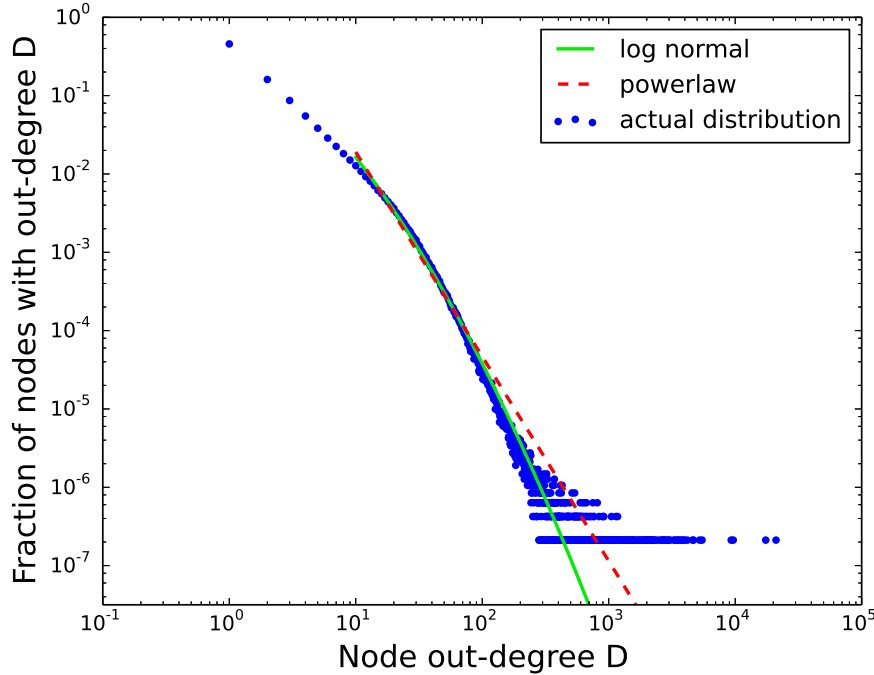


Figure 6.2: Comparison of the best log-normal and power-law fits with  $D_{min} = 10$

## 6.3 Network Evolution

### 6.3.1 Densification

Conventional wisdom about scale-free networks is that the node degrees become bounded. That is to say,  $|E| = \Theta(|V|)$ . However, in [50] the authors find that the network densifies over time, and that the number of edges grows super-linearly like  $\Theta(|V|^{1+c})$ . We find that densification is also present in the Duolingo social graph; however, it does not uniformly densify at all stages of growth. In particular, we find that the graph first becomes *sparser* before densifying. For each day in the evolution of the network we count the total number of nodes present in the graph as well as the number of edges. Figure 6.3 shows the relationship between the number of edges and number of nodes as the network evolves over time.

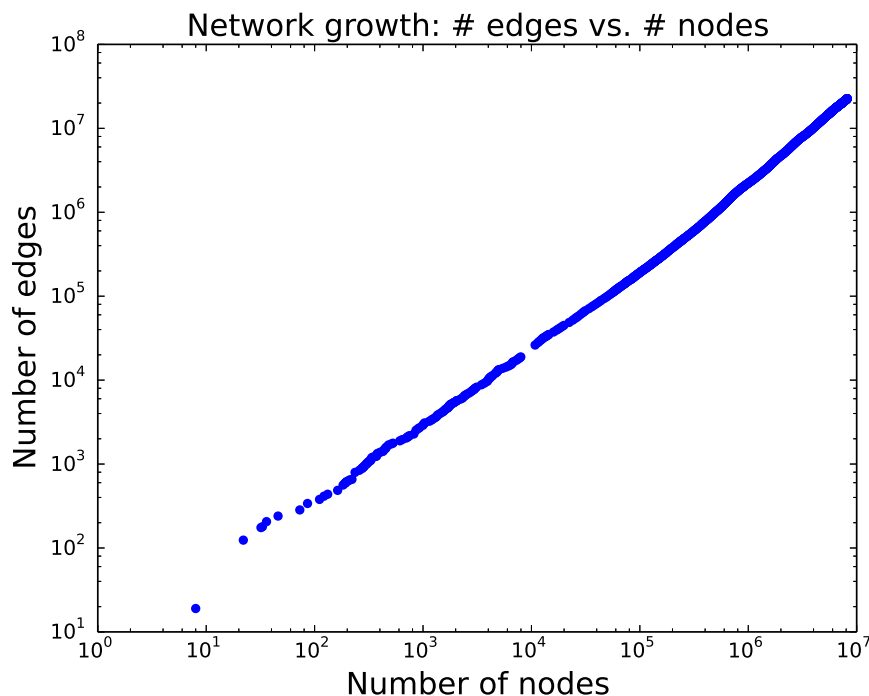


Figure 6.3: Comparing the number of edges vs. the number of nodes in the network for each day since December 1, 2011.

We fit this curve with SciPy [78] using a least-squares power-curve model and find that the fit is quite poor, especially for when the graph is small. This discovery is very perplexing; looking at the average degree vs. network size makes this even more apparent. In figure 6.4 we plot the average out-degree (number of edges divided by number of nodes) and see that the average degree indeed decrease first and then increases once the network reaches approximately 10,000 users.

Without additional context these observations are contradictory to established observations in many other networks. Recall that Duolingo had a private beta period from December 1, 2011 to June 19, 2012 (201 days total). If we look at the network for the first 200 days and from the 201st day onward, we notice something interesting. First, both the linear regression method and the power curve fitting method give very similar results in both cases.

For the first 200 days we find that the scaling exponent is 0.899475 (sparsifying), and after 200 days the scaling exponent is 1.101765 (densifying). One explanation for this is that during the private beta users were invited to the service and would likely not know any other users. After Duolingo opened to the public users could start inviting others in an unrestricted manner

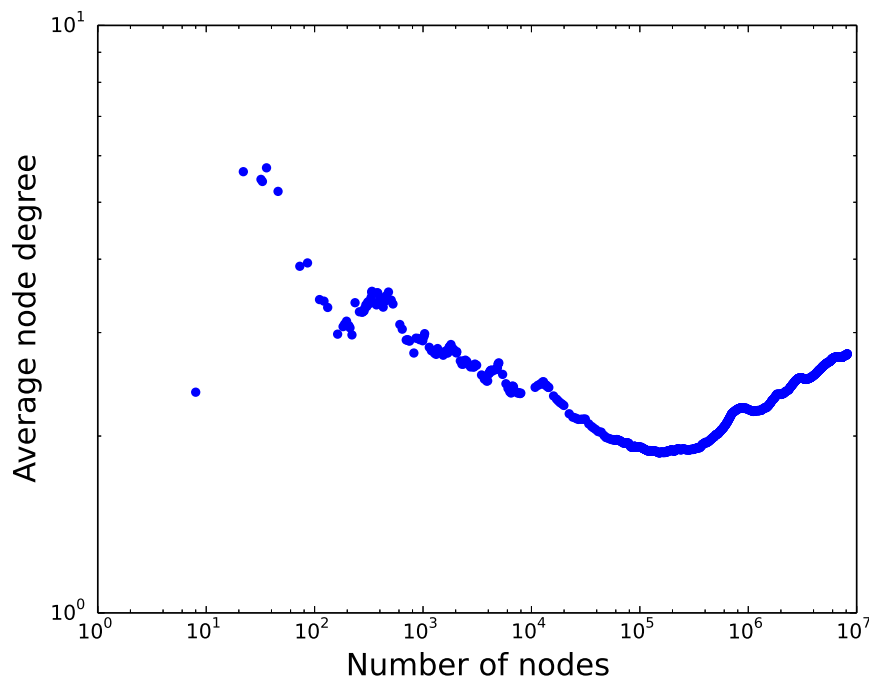


Figure 6.4: Average out-degree vs. number of nodes for each day since December 1, 2011.

and functionality for finding friends who already use Duolingo was implemented.

### 6.3.2 Edge creation delay

We look at the delay between when a student creates their account and when they follow other students. The latency is considered at different timescales of minutes, hours, and days. Immediately, we notice some surprising facts:

- 10,244,376 (45%) of all social connections are created *within one hour* of a student signing up for Duolingo.
- More than half (11,756,788) of the edges in the network are formed within one day of a student creating their account.
- There is a strong cyclic nature to the hourly latency distribution.
- Students continue to follow each other in large numbers long after account creation.

In Figure 6.5 we plot the number of follow events with an  $H$  hour latency, for the first week after a student creates her account. Note that the y-axis is shown on a log scale, otherwise the 0

hour latency count would dominate all other data points. Similar to what we observed in Twitter, there are peaks every 24 hours. One possible explanation for this is that users return to the service around the same time every day. However, we also send reminders to students that they should return to Duolingo. We originally scheduled reminders to be sent at 5:00 pm in the student’s local timezone. After running an A/B test on the time of day to send reminders, we determined that sending reminders in the same hour of the day that a user created their account performs best. To tease apart the effect of receiving reminders versus time-of-day usage patterns we can run an A/B test that offsets the reminder time between the two groups. If the reminder is primarily responsible for this cyclic behavior we should be able to control the peaks in this distribution.

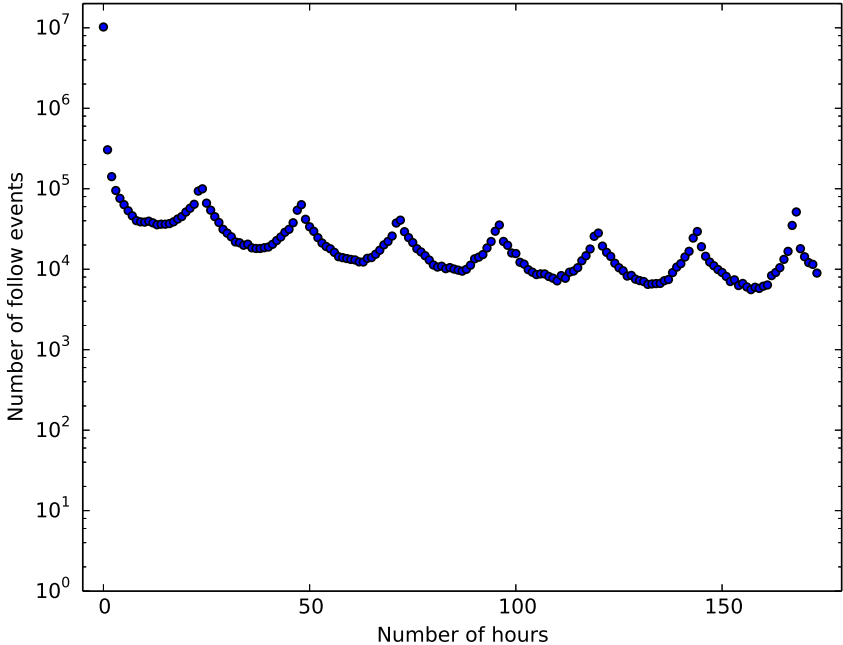


Figure 6.5: The number of edges created  $H$  hours after account creation.

In Figure 6.6 we look at the number of follow events that happen  $d$  days after account creation. Even at very long latencies of more than six months the edge creation rate remains between 1,000 and 10,000 per day. The latency distributions, together with the overall number of edges in the network that come from Facebook connected users, suggest that Facebook is very important to the creation of new edges in the network.

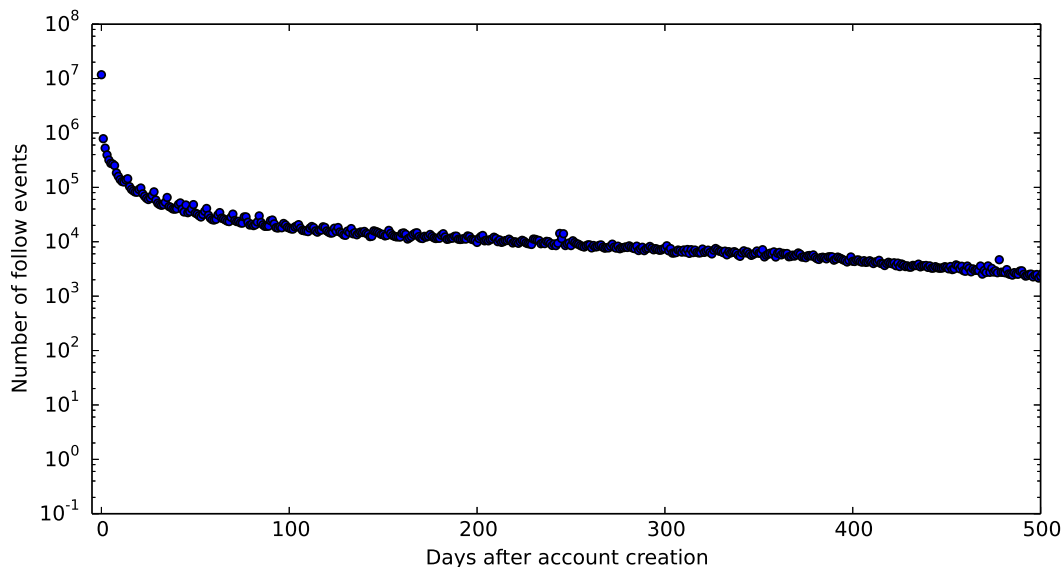


Figure 6.6: The number of edges created  $d$  days after account creation.

Another way to look at the evolution of the graph over time is to compare the number of edges created on a given day to the number of new accounts created on that day. We've seen that half of the edges in the network are formed within 24 hours from when a student created their account. However, this perspective is skewed as we restrict ourselves to looking at students *who end up following other students*. How does the rest of the population behave over time? We look at the number of new edges on a particular day compared to the number of new accounts created on that day in Figure 6.7.

Overall there is a correlation between these two quantities: as the number of new accounts increases so does the number of new edges. It appears as though there is a cluster of points between 150,000 and 200,000 accounts per day. Recall the trial users. On these days we had more traffic from the app stores (possibly due to being featured), and this significantly increased the number of trial user account. In figure 6.8 we only count students that complete the registration process. Now the correlation is much stronger. Being featured in the app stores explains the remaining outliers: the days on which we have gained the most number of new users are when we get featured by Apple or Google. Many users try out featured apps but do not become long-term users (this is confirmed by a drop in daily retention during features). Based on this data it also seems that they are less likely to associate their Facebook account with their Duolingo account.

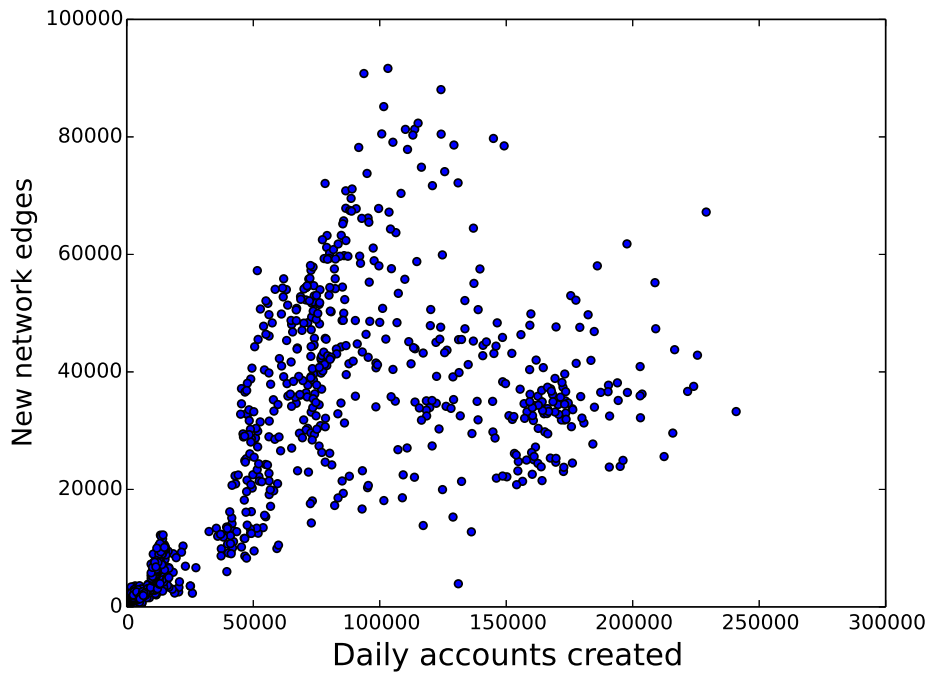


Figure 6.7: New edges vs. daily accounts created

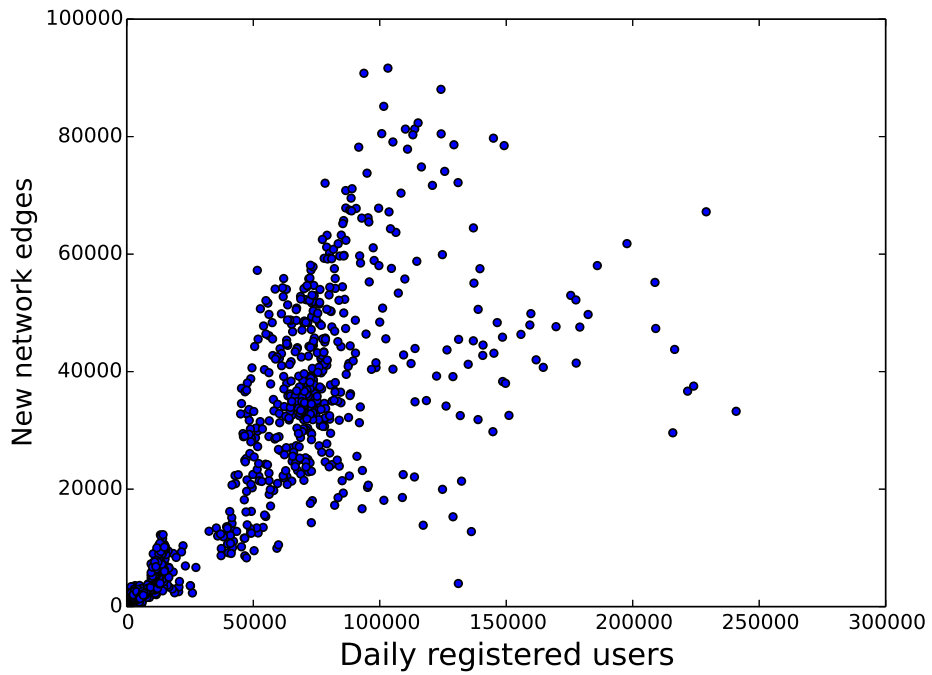


Figure 6.8: New edges vs. daily registrations



### 6.3.3 Other Following Behaviors

We look at two more following behaviors: the delay that it takes for a reciprocal edge to form and the rate at which students follow when we have a ‘follow Duolingo staff’ feature. By default Duolingo sends an email when a student gains a new follower. We have seen other places in Duolingo, such as practice reminders, where emails feature a call-to-action. Looking at the time it takes for a reciprocated edge to form strongly suggests that the new follower emails have a strong impact. In figure 6.9 we plot the fraction of reciprocated edges that form within  $H$  hours of the creation of the initial edge. We find that nearly 40% of reciprocated edges form within *one*

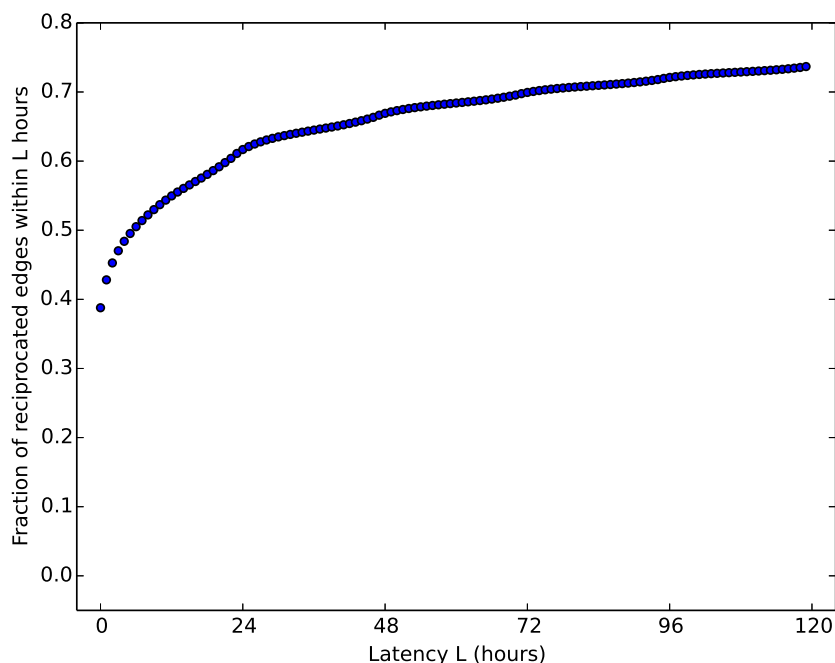


Figure 6.9: Reciprocated edge latency. We plot the fraction of reciprocated edges that form within  $L$  hours of the initial edge forming.

*hour* of the initial edge, and over 60% form within the first day. It is possible to run an experiment in which we artificially introduce a delay in sending the new follower emails to distill the effect of the email from users seeing the notification when coming to the site. Additionally, it is possible to run an A/B test in which we simply do not send new follower emails to discern what impact the notifications have on the creation of reciprocal edges.

Duolingo ran an experiment similar to the suggested users feature in Twitter in which six of

its staff were featured in a step of the account creation process. This experiment had a significant impact on the growth of the network, and somewhat surprisingly, this effect was observed many days after the experiment stopped. In figure 6.10 we plot the number of new edges created each hour in the days leading up to and following the end of the experiment. The experiment started at midnight GMT on May 11 and ran for approximately 24 hours.

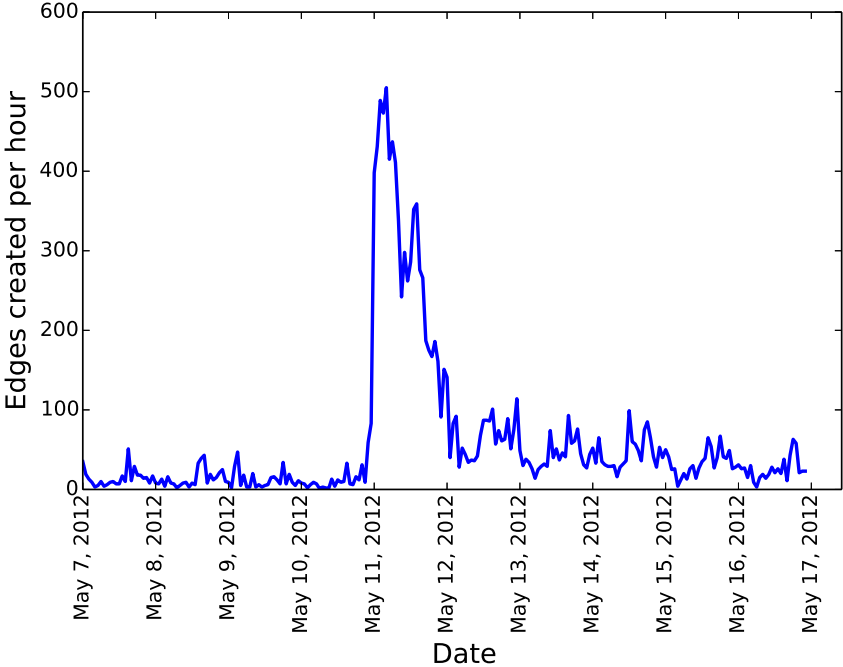


Figure 6.10: The hourly rate at which new edges are created during the suggested user experiment.

While the experiment was running new edges were created almost twenty times more often than before the experiment started. Although the rate decreased substantially after the experiment ended there is still a clearly sustained impact on user following behavior. One possible explanation for this continued increase in follow behavior is that the suggested user prompt simply made new students aware of the social features. Students exposed to the prompt who return to the site several days after signing up may be more likely to search for their friends or follow other students in the forums. Overall, this experiment showed the powerful impact product decisions have on user behavior and network evolution and highlights the importance of having a thorough understanding of product changes when studying graph dynamics in online social networks.

# Chapter 7

## Student Behavior in Duolingo

Thus far we have presented an overview of the Duolingo ecosystem and the structure and evolution of its social network. In this chapter we examine how user behavior changes based on social interactions and motivators such as the streak.

### 7.1 The Impact of Using Social Features

This analysis focuses on a group of students who signed up between May 28, 2014 and June 11, 2014. In this time one million students created accounts and serve as the cohort we study for understanding the impact of social feature use on student progress and retention. Within this cohort we look at several populations based on which social features they use. We first study whether attaching one's Duolingo account to one's Facebook account (thereby enabling a student to easily find friends who use Duolingo) impacts the number of lessons done. Next, we see the effect of following other students and being followed. Finally, we take into account the time over which a user interacts with the social features. A student who starts following another student many days after creating their account is necessarily considered active after that long period (as they took some action on Duolingo). We control for this by considering students who create all of their social connections within one day or one week after creating their account.

### 7.1.1 Impact on Progress

We first look at how using different social features impacts the number of lessons done by a student on Duolingo. The number of lessons completed is a good proxy for the amount of progress a student makes; the overwhelming majority of students do not place out of lessons when they sign up for the service. For each of the sub-populations (as determined by use of social features) we bucket students based on the number of lessons they completed. The distribution of the number of lessons completed is shown in figure 7.1. A lognormal distribution with parameters  $\mu = 1.72829461506$ ,  $\sigma = 1.49053268712$  fits the empirical data extremely well.

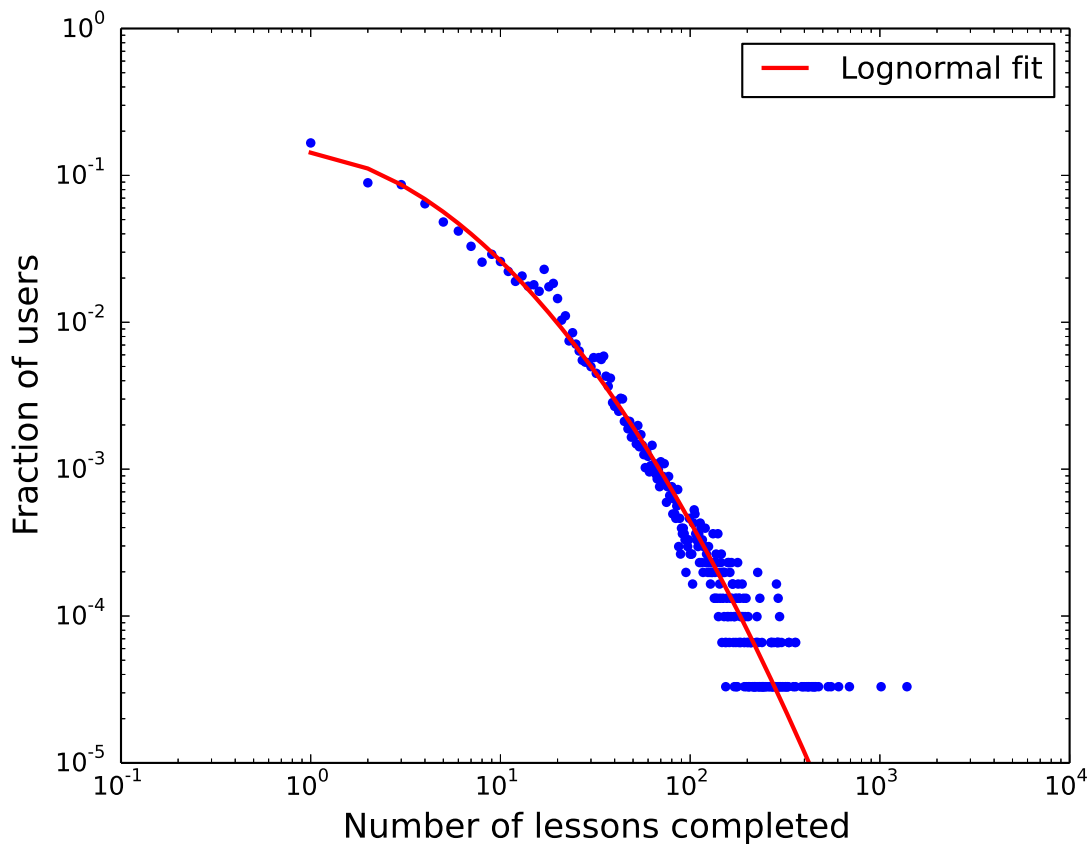


Figure 7.1: The empirical and fitted lognormal distribution for the number of lessons completed.

First we consider the impact of connecting one's account to Facebook. In figure 7.2 the histogram of lessons completed for each of the three populations is given. The number of lessons completed is bucketed exponentially (by powers of 2) to group students more coarsely. We find that students connecting with Facebook tend to do slightly more lessons, with fewer of

them (21% vs. 25% in the overall population) doing exactly zero lessons. The subpopulation of students who do not connect their account to Facebook is not noticeably different from the population at large. This is consistent with the fact that most students do not connect their accounts to Facebook.

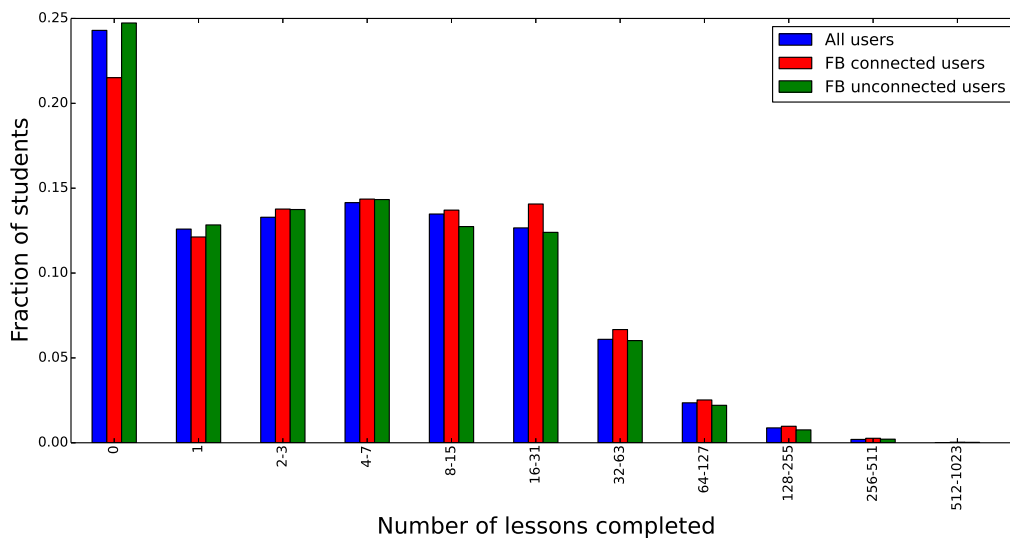


Figure 7.2: Histogram of number of lessons completed based on connecting one's account to Facebook.

We next consider the impact of following other students or being followed by other students. It is possible to connect one's account to Facebook but not have any friends on Duolingo. However, once a student follows other students they have a non-trivial leaderboard in which they can see weekly, monthly, and all-time progress. As seen in figure 7.3, students who follow other students or are themselves followed tend to do *significantly* more lessons than a sample of 40,000 students in the overall population. This difference is far more pronounced than the impact by connecting or not connecting a Duolingo account to Facebook. Over half of the students who either follow or are followed complete at least 16 lessons, compared to only 29% of users from the overall population.

Finally, we control for the the time in which students use the social features on Duolingo. We consider (not necessarily disjoint) are those users who start following others within the first day of creating their account. These students are the early-adopters of social features. Additionally,

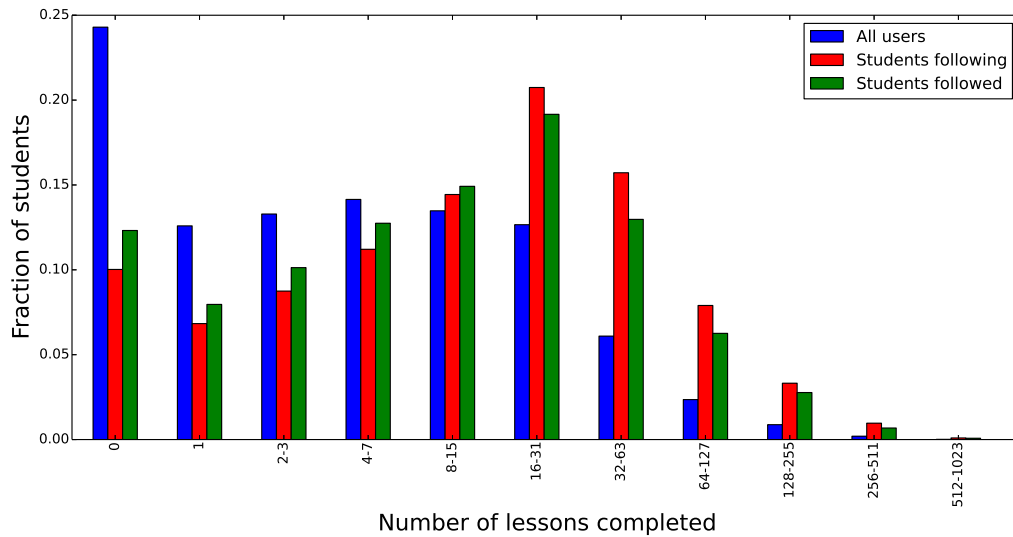


Figure 7.3: Histogram of number of lessons completed based on following or being followed.

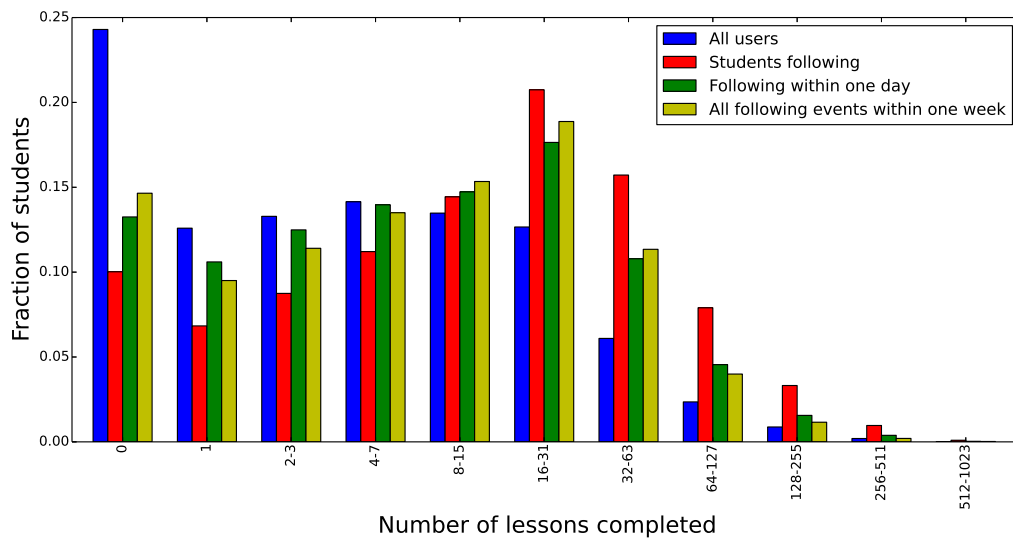


Figure 7.4: Histogram of number of lessons completed based on different following conditions.

we consider the set of students who finished creating all of their social connections within one week of creating their account. Bounding this time period will be more interesting in the next section when we examine how long students use Duolingo. As we see in figure 7.4, both of

these subpopulations do more lessons than the population as a whole, but do fewer lessons than a sample of students who follow others but have no time restrictions on their following behavior.

### 7.1.2 Impact on Student Lifetime

Now that we have seen the impact of different social components on the progress a student makes, we want to see whether student lifetime (how long a student is active on Duolingo) is impacted similarly. For the sake of this analysis we look at when a student is active in any capacity, and not just completing lessons. We then look at what fraction of each population is active at least  $D$  days after registering. This measure is similar to the complementary cumulative density function if we were to consider the last day on which a student is active to be a probability density.

Figure 7.5 shows that connecting one’s Duolingo account to Facebook only slightly increases retention, and not connecting only slightly decreases it. For lifetimes  $L$  of 20 days or longer there is at most a 5% absolute increase in the fraction of students connected to Facebook with lifetime at least  $L$  compared to the general population.

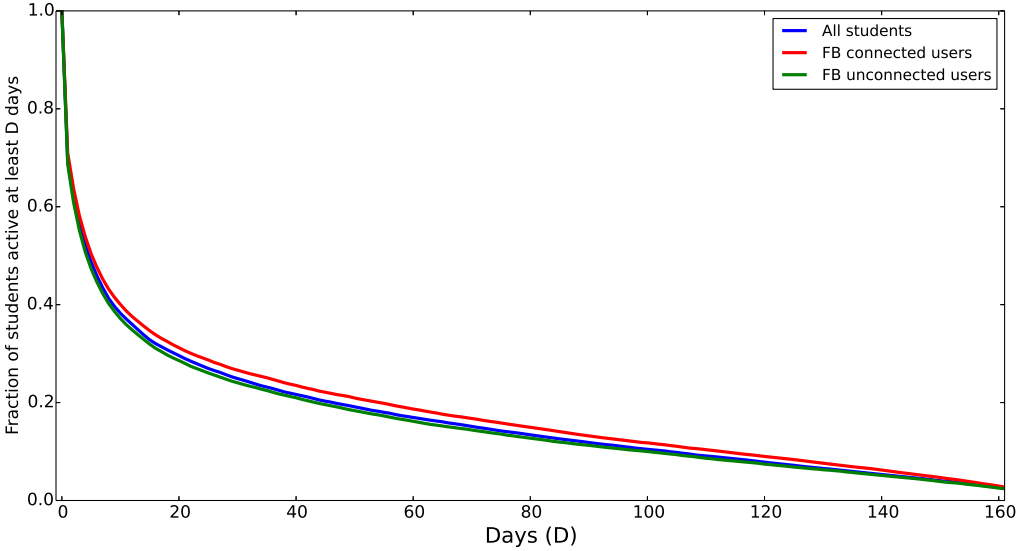


Figure 7.5: Student lifetime based on Facebook connectivity.

A much more significant impact in lifetime is seen when looking at students who follow other students. Compared to the overall population, students who follow or are followed have far longer lifetimes, as seen in figure 7.6.

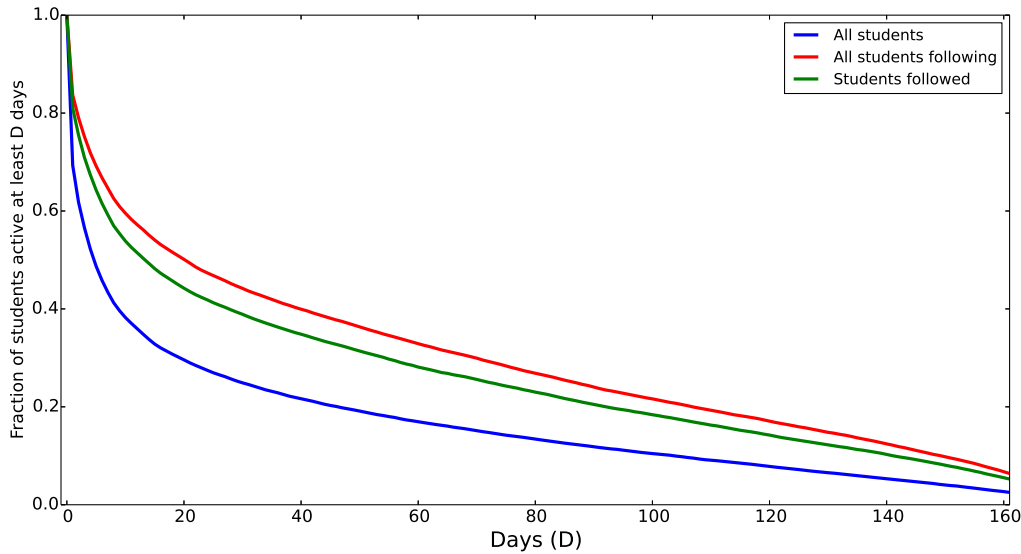


Figure 7.6: Student lifetimes based on following or being followed.

However, this brings us back to our original point of following another student necessarily increasing lifetime. If a student  $S$  starts following another student 100 days after creating their account, then  $S$ 's lifetime is at least 100 days. We control for the time after account creation in which the follow events happen (thereby limiting their ability to increase student lifetime) and see that *following alone does not significantly increase student lifetime!* This is very surprising given that this population of students did significantly more lessons than the overall population. Looking into this matter further shows that these students do much more shortly after creating their accounts, but that their activity tapers off rapidly thereafter. The student lifetimes for the conditions of following with one day and making all following connections within one week are given in figure 7.7. We observe that when time restrictions are in place that the lifetime for users does not significantly increase. Somehow students who use the social features early on do significantly more lessons but do not stay around significantly longer than their peers.

We also looked at whether the student lifetime distribution fits a power-law or some other heavy-tailed distribution and find that neither the head nor the tail of the distribution fits a power-law or log-normal distribution well. The head of the distribution decays too rapidly and the tail too slowly for either to provide a good fit.



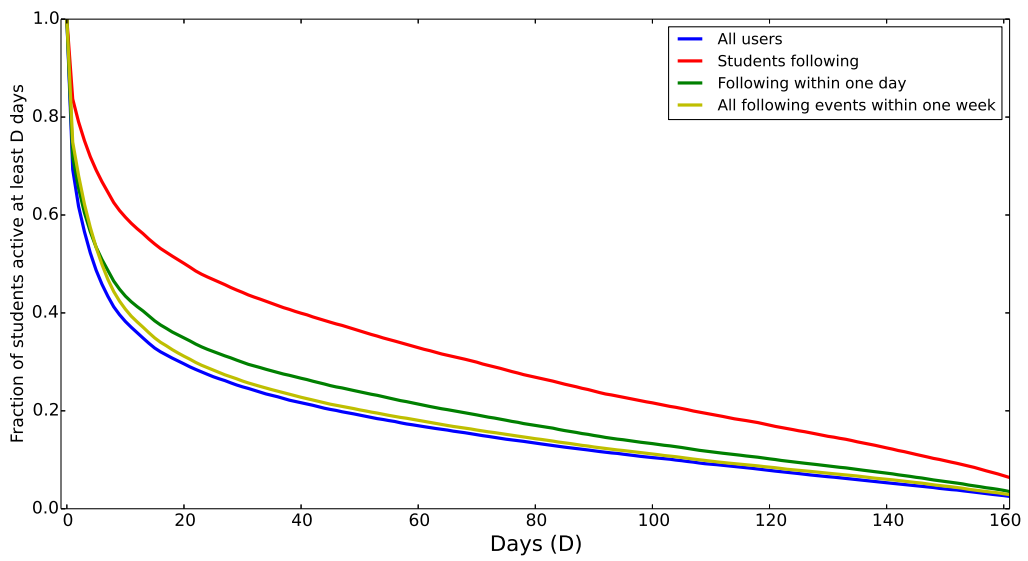


Figure 7.7: Student lifetime for students following, with different time requirements for when the follow events occur.

## 7.2 Streaks and Continued Use

Among the various gamification elements in Duolingo, the coach and streak have a pronounced impact on student retention and long-term use. The fundamental question we want to answer is: “What is the probability that a student extends (increases) her  $D$  day streak?” If this probability increases as the streak length increases it would suggest that the streak and coach are strong motivators for continued student engagement. Whenever a student loads their home screen (either by visiting the website or opening the app) we log their current streak.

Looking at the streak logs over consecutive days enables us to empirically calculate the probability that a student with a  $D$  day streak will reach a  $D + 1$  day streak. We call this probability the *continuation probability* for a streak of length  $D$ , and is calculated by the fraction of students who reach a streak of length  $D$  that also reach a streak of length  $D + 1$ . Using these logs from March 1, 2015 to March 10, 2015 we calculate the continuation probabilities. As seen in figure 7.8, the continuation probability increases significantly as the streak length goes from 0 days to 20 days. After 30 days there is little change in the continuation probability, suggesting that the motivating effect of the streak reaches a maximum after the one-month milestone.

The distribution of streak lengths is heavy-tailed, as seen in figure 7.9, with the best-fit power-law curve having scaling exponent  $\alpha = -1.7670$  and  $D_{min} = 6$ .

### 7.2.1 Streak Length and Platform Usage

Given that the streak feature is implemented in different ways, and that the features of the mobile apps differ from the website, we expect there to be some variation in platform use. Specifically, we expect that the fraction of users with short streaks (zero days or one day) closely follows the breakdown of the platform for new users. We hypothesize that students with longer streaks will mostly be using the website or the website and some mobile platform together. The main reason we believe this to be the case is that the increased number of ways to earn XP on the website, in addition to the timed practice feature, allows for students to quickly come to Duolingo and extend their streak. Additionally, the total streak length is very prominently displayed in a student’s home screen, and appears as part of the flair in the discussion.

In Figure 7.10 and Figure 7.11 we look at the distribution of platforms among students with a particular streak length at the day and week granularity. The proportions are reported by taking

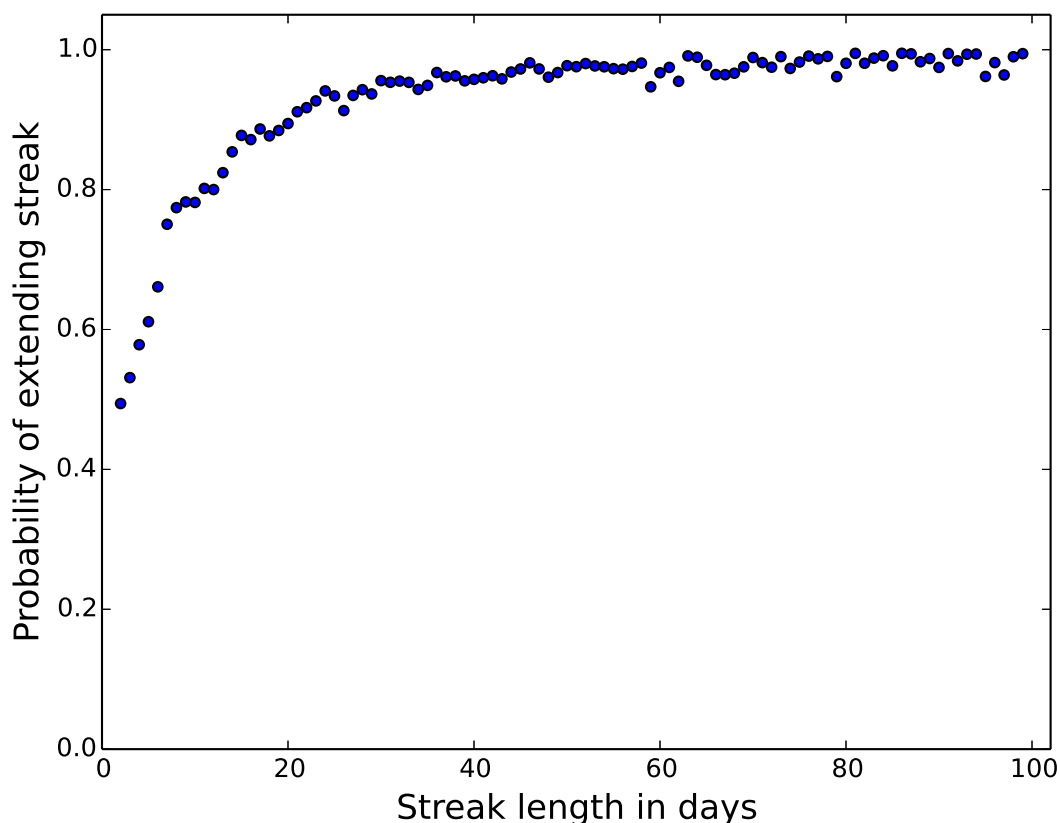


Figure 7.8: Empirical continuation probabilities for users in the first ten days of March, 2015.

the number of students with a  $D$  day streak and dividing by the total number of students with a  $D$  day ( $W$  week) streak. A single student using multiple platforms will be counted towards each platform she uses, but we find that only a small fraction of students use multiple platforms in a single day. This data is based on over one million students who used Duolingo on March 5, 2015.

In Figure 7.10 we look at the distribution of platforms among of students with a  $D$  day streak for streaks up to 100 days. Generally the 0-day streak proportions closely follow the breakdown of the student population by platform. We find that as streak lengths increase the proportion of students using the website does indeed increase. There is large variability in the proportions because the number of students having a streak of  $D$  days for  $D > 50$  is relatively small.

In Figure 7.11 we look at the distribution of platforms among of students with a  $W$  week streak for streaks up to 52 weeks (one year). The solid lines convey the proportion, and the

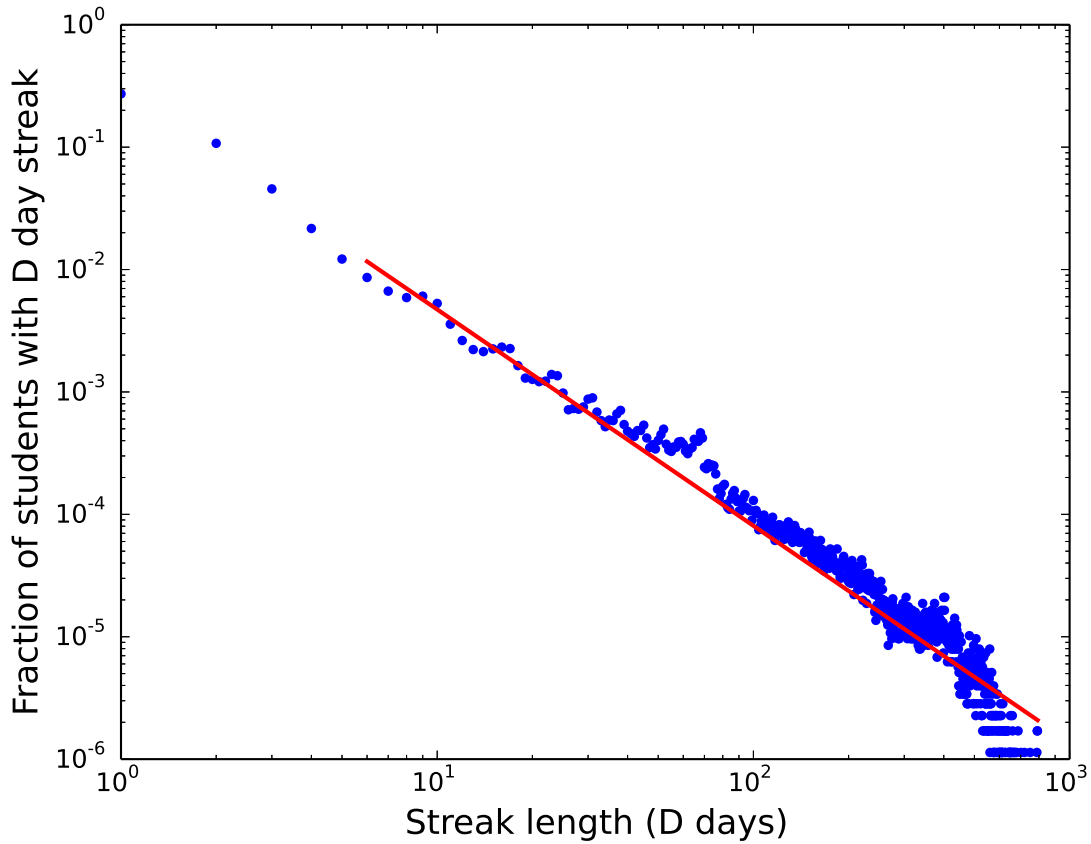


Figure 7.9: Distribution of streak lengths and the best-fit power-law distribution ( $\alpha = -1.7670$ ,  $D_{min} = 6$ )

dashed lines show the 95th percentile multinomial confidence interval. We see that more than half of students with a streak 15 weeks or longer use the website. There is no statistically-significant difference between the two major mobile platforms, although iOS almost always has a higher proportion of students.

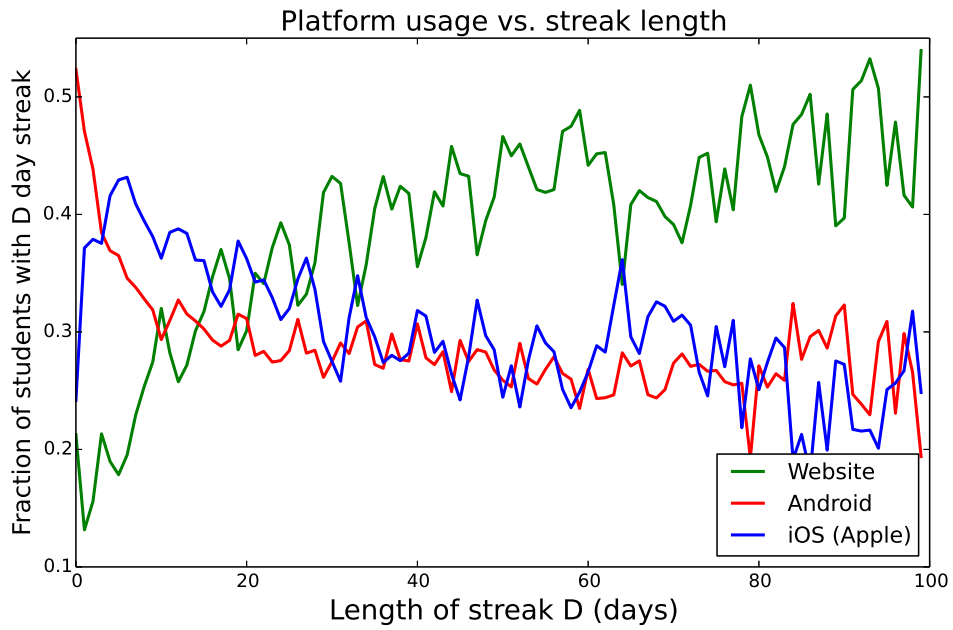


Figure 7.10: Proportions of students per platform, bucketed by streak length in days.

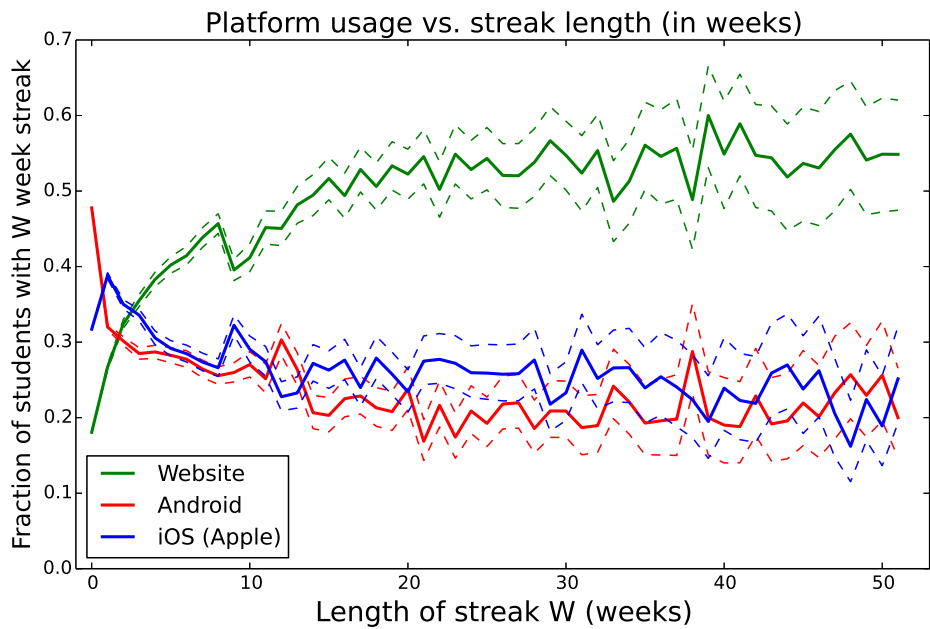


Figure 7.11: Proportions of students per platform, bucketed by streak length in weeks. The dotted lines mark the multinomial 95th percentile ranges for each proportion.

In conclusion, we find that the streak is an effective motivator and indicator of continued use on Duolingo. Once students reach a sufficiently long streak they will continue to pursue extending their streak with high probability. We see a shift in platform use when segmenting users by streak length; the longer a user's streak, the more likely it is that they use the website or multiple platforms. Losing a streak does have a tendency to cause students to abandon Duolingo, and this impact is more pronounced as the length of the streak increases. However, the benefit of users being likely to continue to use Duolingo because of reaching a certain streak outweighs the cost of user abandonment due to losing one's streak.

# Chapter 8

## Conclusion

### 8.1 Topical Differences in Information Diffusion

By studying the ways in which an individual's use of widely-adopted Twitter hashtags depends on the usage patterns of their network neighbors, we have found that hashtags of different types and topics exhibit different mechanics of spread. These differences can be analyzed in terms of the probabilities that users adopt hashtags after repeated exposure, with variations occurring not just in the absolute magnitudes of these probabilities but also in their rate of decay. Some of the most significant differences in hashtag adoption provide intriguing confirmation of sociological theories developed in the off-line world. In particular, the adoption of politically controversial hashtags is especially affected by multiple repeated exposures, while such repeated exposures have a much less important effect on the adoption of conversational idioms.

This extension of information diffusion analysis, or taking into account sources of variation across topics, opens up a variety of further directions for investigation. First, the process of diffusion is well-known to be governed both by influence and also by homophily — people who are linked tend to share attributes that promote similarities in behavior. Recent work has investigated this interplay of influence and homophily in the spreading of on-line behaviors [7, 9, 20, 40]; It would be interesting to look at how this varies across topics and categories of information as well — it is plausible, for example, that the joint mention of a political hashtag provides stronger evidence of user-to-user similarity than the analogous joint mention of hashtags on other topics, or that certain conversational idioms (those that are indicative of shared background) are significantly better indicators of similarity than others. There has also been work on the temporal

patterns of information diffusion — the rate over time at which different pieces of information are adopted [21, 36, 47, 55, 87]. In this context there have been comparisons between the temporal patterns of expected versus unexpected information [21] and between different media such as news sources and blogs [47]. Our analysis here suggests that a rich spectrum of differences may exist across topics as well.

Finally, we should emphasize one of our original points, that the phenomena we are observing are clearly taking place in aggregate: it is striking that, despite the many different styles in which people use a medium like Twitter, sociological principles such as the complex contagion of controversial topics can still be observed at the population level.

## 8.2 Network Evolution in a Large Subgraph

We have devised a simple and effective method for inferring follow times in the Twitter social network that has several distinct advantages over other ways of recovering this information. We are able to accurately and robustly infer link creation times using only a single crawl of the social network and user creation times. Furthermore, we are able to recover follow times arbitrarily far into Twitter’s history. For the most popular users in Twitter’s social network, the method was accurate to within several minutes.

Using the timestamp information, we recreated the evolution of the Twitter celebrity subgraph and gained temporal insights into user following behavior including the distribution of latencies, the importance of the Twitter interface, and the possible influence of real-world events. Overall, our approach gave us a much deeper insight into the structure and evolution of a significant and large subgraph of the Twitter social network.

We have also examined the growth of the social network in Duolingo. By encouraging students to connect Facebook with their Duolingo account we are able to bootstrap a large number of social edges. The majority of edges in the network are created very shortly after students create their accounts. At long times we see that many new edges are added from users hundreds of days after they start using Duolingo. Many of the connections in the network are between users in the same language.



## 8.3 Social Networks in Online Education Platforms

We present findings about how students use Duolingo over long periods of time. Engaging students with reminders and other nudges greatly increases the chances of them returning to the service. The social features in Duolingo are used by a non-negligible fraction of students and correlate with prolonged use and increased learning activity. We have reason to believe that our notifications have a causal impact on student behavior. Furthermore, previous experiments with social features, such as removing the leaderboard, have demonstrated their positive impact on student retention. Because we deeply value the user experience, we must devise experiments that can be run for long durations of time without negatively affecting students. Removing existing features for new users, only to find that their presence really is important, is not a viable way for us to run experiments. Instead, we must explore the space of product improvements and additions. One proposal is to show students a notification to revisit their Facebook connections. We've seen the initial connection to Facebook results in a very large number of edges in the social network. With our massive growth in student population over the last year, older users would likely discover their friends are now using Duolingo. Developing entirely new social features is in the realm of possibilities. We tried one such feature, called *Duels*, that allowed students to practice their vocabulary in a competitive manner. This feature was eventually removed as it was not effective at increasing user engagement. One possible route is to provide a *groups* or *teams* feature in which students have greater visibility into each other's activity. This could operate similar to clans in many massive online role playing games. Incentives such as competition within a group or between groups could increase retention over long durations.

## 8.4 Future work: Increased Understanding of Student Behavior

One line of work in the future consists of better modeling of student behavior. Building models that can accurately predict whether a user will return (alternatively, whether a user is at risk of *not* returning) enable practitioners to take actions to engage users. This seems especially important for continued user engagement in activities such as language learning or physical training. We also consider building generative models that can reach the same predictive ability. With such

models we may be able to identify different kinds of users, such as those who learn in a bursty manner or those who do a small amount consistently. It is probably the case that the efficacy of reminders, marketing, and motivators varies among the different student populations. The ability to identify and target these populations is an opportunity that can increase user retention and learning outcomes.

## **8.5 A Few Afterthoughts**

I'm extremely proud of the work that the Duolingo team has done. Building, maintaining, and improving a large production system with many millions of users is no easy task. It has been incredibly rewarding to work on a project with such impact. The experience of working on Duolingo has shaped this thesis in a few ways. First, I have a far greater appreciation for the balance between engineering, design, and experimentation. Realizing that 'minor' changes in a user interface can have a far greater impact on user retention than a more accurate way to predict whether a student knows a word is humbling and eye-opening. Duolingo is an example of how outstanding design, experimentation, machine learning, and computer systems come together to provide significant value.

One of the most valuable lessons learned while building Duolingo is that many complications exist in data collection, even when one has full control of a system. Products with significant complexity and large user bases are created by teams of people, not individuals. We have seen how developers can sometimes make changes to each other's logging, directly or indirectly. Several times this has impacted our understanding of an A/B test or our ability to process backend logs. Simply collecting the data is not enough- it must be verified and monitored over time just like the health of any production system. Twitter has written [45] about unified logging and outlines some steps organizations can take to reduce some of the headaches we have experienced.

### **Play nicely with others**

Over the course of eighteen months we used the Twitter API to collect a social graph with over 4.5 billion edges between over 130 million users and over 17 billion tweets. This was done using a distributed crawler running on eighty machines at Carnegie Mellon. Based on publicly released figures about Twitter's traffic when this collection was occurring, the crawler attributed for 1-2% of their traffic. The crawler lacked sophistication such as exponential back-off for

when failures occur. At one point Twitter experienced site stability issues while the crawler was running; this certainly did not help the stability issue. As a result a block of CMU IP addresses were temporarily banned from accessing Twitter; this matter was quickly resolved. Until at least 2013 this was the most comprehensive collection of data available outside of Twitter in an academic institution. Today, Twitter has arrangements [85] to provide data to organizations and institutions so that researchers can focus on addressing their questions without abusing Twitter's compute and bandwidth resources. I think that this is a positive direction that benefits both Twitter and the research community, and hope that other companies do the same in the future.



# Bibliography

- [1] About Twitter API. <http://dev.twitter.com>. 3.1
- [2] L. Adamic. Friends and Neighbors on the Web. *Social Networks*, jan 2003. 4.3.2
- [3] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004. 2.1
- [4] L. Akoglu, H. Tong, B. Meeder, and C. Faloutsos. Pics: Parameter-free identification of cohesive subgroups in large attributed graphs. In *SDM*, pages 439–450. Citeseer, 2012. 1.2
- [5] J. Alstott, E. Bullmore, and D. Plenz. powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS One*, 9(1):e85777, 2014. 6.2.1
- [6] Amazon web services (aws) cloud computing services. <http://aws.amazon.com>. Accessed: 2014-11-28. 5.6
- [7] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–15, 2008. 8.1
- [8] Anil Dash’s experience on the suggested users list. <http://dashes.com/anil/2009/12/life-on-the-list.html>. 3.3.2
- [9] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA*, 106(51):21544–21549, Dec. 2009. 8.1
- [10] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006. 2.1
- [11] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of

the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, Aug. 2002. 4.3.2

- [12] S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28:124–136, 2006. 3.1
- [13] C. Borgs, J. Chayes, B. Karrer, B. Meeder, R. Ravi, R. Reagans, and A. Sayedi. Game-theoretic models of information overload in social networks. In *Algorithms and Models for the Web-Graph*, pages 146–161. Springer, 2010. 1.2
- [14] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 3 September 2010. 2.1, 2.1
- [15] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113:702–734, 2007. 2.1, 2.1, 2.4
- [16] J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56. IEEE, 2011. 1.2, 0
- [17] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. 3.1
- [18] D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *Proc. 4th International Conference on Weblogs and Social Media*, 2010. 2.1, 2.1, 2.1, 2.3
- [19] D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, X. Lan, and S. Suri. Sequential influence models in social networks. In *ICWSM*, 2010. 3.1
- [20] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 160–168, 2008. 8.1
- [21] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA*, 105(41):15649–15653, 29 September 2008. 2.1, 8.1

- [22] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079–1187, 2002. 3.1
- [23] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proc. 27th ACM Conference on Human Factors in Computing Systems*, pages 211–220, 2009. 2.4, 4.1.2
- [24] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Little, Brown, 2000. 2.1
- [25] M. Gladwell. Small change: Why the revolution will not be tweeted. *The New Yorker*, 4 October 2010. 2.1
- [26] S. A. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Proceedings of the Second IEEE International Conference on Social Computing*, 2010. 3.1
- [27] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973. 2.1
- [28] M. Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1:201–233, 1983. 2.4
- [29] D. Gruhl, D. Liben-Nowell, R. V. Guha, and A. Tomkins. Information diffusion through blogspace. In *Proc. 13th International World Wide Web Conference*, 2004. 2.1
- [30] S. B. H. Hacker. Duolingo: Learning a language while translating the web. PhD Thesis CMU-CS-14-116, Carnegie Mellon University, 2014. 5
- [31] C. Heath and D. Heath. *Made to Stick: Why Some Ideas Survive and Others Die*. Random House, 2007. 2.1
- [32] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *Social Science Research Network Working Paper Series*, December 2008. 3.1
- [33] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), Jan. 2009. 1
- [34] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM. 3.1

- [35] H. Jeong, Z. Néda, and A. L. Barabási. Measuring preferential attachment in evolving networks. *Europhysics Letters*, 61(4):567–572, 2003. 3.1
- [36] A. Johansen. Probing human response times. *Physica A*, 338(1–2):286–291, 2004. 8.1
- [37] U. Kang, B. Meeder, and C. Faloutsos. Spectral analysis for billion-scale graphs: Discoveries and implementation. In *Advances in Knowledge Discovery and Data Mining*, pages 13–25. Springer, 2011. 1.2
- [38] U. Kang, B. Meeder, E. E. Papalexakis, and C. Faloutsos. Heigen: Spectral analysis for billion-scale graphs. *Knowledge and Data Engineering, IEEE Transactions on*, 26(2):350–362, 2014. 1.2
- [39] L. Katz. A New Status Index Derived From Sociometric Analysis. *Psychometrika*, 1953. 4.3.2
- [40] G. Kossinets and D. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 115(2):405–50, Sept. 2009. 8.1
- [41] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM. 3.1
- [42] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, 2006. 3.1
- [43] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media. In *WWW'10: Proceedings of the 19th international conference on World Wide Web*, pages 591–600, 2010. 3.1, 3.3.1
- [44] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *Proc. 19th International World Wide Web Conference*, pages 591–600, 2010. 4.1, 4.1.2
- [45] G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy. The unified logging infrastructure for data analytics at twitter. *Proceedings of the VLDB Endowment*, 5(12):1771–1780, 2012. 8.5
- [46] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1), May 2007. 2.1



- [47] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009. 8.1
- [48] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM. 3.1
- [49] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proc. 28th ACM Conference on Human Factors in Computing Systems*, pages 1361–1370, 2010. 4.1.2
- [50] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005. 6.3.1
- [51] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007. 3.1
- [52] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. SIAM International Conference on Data Mining*, 2007. 2.1
- [53] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007. 4.1.2
- [54] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA*, 105(12):4633–4638, Mar. 2008. 2.1
- [55] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci. USA*, 105(47):18153–18158, 25 November 2008. 8.1
- [56] D. McAdam. Recruitment to high-risk activism: The case of Freedom Summer. *American Journal of Sociology*, 92:64–90, 1986. 2.1
- [57] D. McAdam. *Freedom Summer*. Oxford University Press, 1988. 2.1

- [58] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on World wide web*, pages 517–526. ACM, 2011. 1.2, 0
- [59] B. Meeder, J. Tam, P. G. Kelley, and L. F. Cranor. Rt@ iwantprivacy: Widespread violation of privacy settings in the twitter social network. In *Proceedings of the Web*, volume 2, 2010. 1.2
- [60] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42, New York, NY, USA, 2007. ACM. 3.1
- [61] Mixpanel — mobile analytics. <http://www.mixpanel.com>. Accessed: 2014-11-28. 5.5
- [62] M. Motoyama, B. Meeder, K. Levchenko, G. M. Voelker, and S. Savage. Measuring online service availability using twitter. *WOSN'10*, pages 13–13, 2010. 1.2
- [63] MTV Video Music Awards 2009. <http://www.mtv.com/ontv/vma/2009/>. 3.3.4
- [64] National center for education statistics: Fast facts. <http://nces.ed.gov/fastfacts/display.asp?id=372>. Accessed: 2014-11-26. 1.1
- [65] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, (2):025102+, July 2001. 4.3.2
- [66] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003. 3.1
- [67] B. O'Connor, R. Balasubramanian, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010. 1.2
- [68] Oprah Tries Twitter, Crowns Ashton King of It. <http://blogs.wsj.com/digits/2009/04/17/oprah-tries-twitter-crowns-ashton-king-of-it/>. 3.3.4
- [69] Powerlaw: A toolbox for testing if a distribution fits a power law (software library). <https://pypi.python.org/pypi/powerlaw>. 6.2.1
- [70] S. Redner. Citation statistics from 110 years of Physical Review. *Physics Today*, 58:49–54,

2005. 3.1

- [71] E. Rogers. *Diffusion of Innovations*. Free Press, fourth edition, 1995. 2.1
- [72] D. Romero and J. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proc. 4th International AAAI Conference on Weblogs and Social Media*, 2010. 3.1, 3.2
- [73] D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proc. 4th International Conference on Weblogs and Social Media*, 2010. 4.1, 4.1.2
- [74] D. M. Romero, B. Meeder, V. Barash, and J. M. Kleinberg. Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness. In *ICWSM*, 2011. 1.2
- [75] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011. 1.2, 0
- [76] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, 2011. 4.2.2
- [77] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. 4.3.2
- [78] Scipy: Scientific computing in python. <http://www.scipy.org/>. 6.3.1
- [79] D. Strang and S. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998. 2.1
- [80] E. Sun, I. Rosenn, C. Marlow, and T. M. Lento. Gesundheit! Modeling contagion through Facebook News Feed. In *Proc. 3rd International Conference on Weblogs and Social Media*, 2009. 2.1
- [81] The Newest Way to Game Twitter Fake Followers. <http://brooksbayne.com/post/79132853/the-newest-way-to-game-twitter-fake-followers#>

comment-6353220. 3.3.1

- [82] Twitter blog post: Discovering Who To Follow. <http://blog.twitter.com/2010/07/discovering-who-to-follow.html>. 3.3.1
- [83] Twitter blog post: Suggested Users. <http://blog.twitter.com/2009/03/suggested-users.html>. 3.3.1
- [84] Twitter blog post: The Power of Suggestion. <http://blog.twitter.com/2010/01/power-of-suggestions.html>. 3.3.1
- [85] Introducing twitter data grants. <https://blog.twitter.com/2014/introducing-twitter-data-grants>. Accessed: 2015-04-20. 8.5
- [86] J. R. Tyler and J. C. Tang. When Can I Expect an Email Response? A Study of Rhythms in Email Usage. In *ECSCW'03: Proceedings of the eighth conference on European Conference on Computer Supported Cooperative Work*. Kluwer Academic Publishers, Sept. 2003. 4.1.2
- [87] A. Vazquez, J. G. Oliveira, Z. Deszo, K.-I. Goh, I. Kondor, and A.-L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(036127), 2006. 8.1
- [88] Who Does Twitter Love? Breaking Down The Twitter Suggested Users List. <http://searchengineland.com/who-does-twitter-love-breaking-down-the-twitter-suggested-users-list-22640>. 3.3.1