Computational methodologies for understanding the dynamics of an online community of educators

by

Ogheneovo Dibie

BS., University of Maine, 2010
M.S., University of Colorado Boulder, 2012

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
2016
This thesis entitled:
Computational methodologies for understanding the dynamics of an online community of educators
written by Ogheneovo Dibie
has been approved for the Department of Computer Science

Prof. Tamara Sumner

Prof. James Martin

Dr. Keith Mauull

Date ________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Online learning communities are becoming an invaluable component of educator instruction. By providing educators with access to teaching resources and best practices shared by their peers, these communities have been shown to improve the instructional practices of educators and produce increases in student learning. Given the importance of online learning communities to teaching and learning, understanding their dynamics and the factors that influence these dynamics has key implications for educator instruction and student learning. A better understanding of the aforementioned dynamics can also benefit agencies that support these communities.

In this dissertation, I show that sociological network theory can be used to understand the dynamics of online learning communities. Specifically, the phenomena of homophily (tendency of individuals to have social ties with others of similar traits) and triadic closures (tendency of new connections to develop between individuals sharing a common neighbor) can be understood through the sharing and usage behaviors of educators. I also demonstrate how an understanding of the triadic closure process can be used to improve the performance of traditional resource recommendation systems. Finally, I show that social influence may play a significant role in the diffusion and popularity of resources within online learning communities.
Dedication

To all those with an insatiable appetite for learning from data.
Acknowledgements

I would like to begin by thanking my advisor, Tammy. Thank you for taking me on as a starry-eyed fresh college graduate. I am eternally grateful for your guidance, support and patience as I sought and oftentimes struggled to find my research bearing.

I would also like to thank my past and present colleagues. Keith, thank you for being a big brother to me. Your love for data and all things technology fascinates and inspires me. Inasmuch as you love Python, I hope you’ll one day experience the unrivaled delirium of working in R. M.G., thank you for your ever-present guidance and support. I am grateful for your willingness to lend a helping hand whenever I needed one. Soheil, Arafat, David and Ifeyinwa, I am deeply honored to have worked with you over the past few years. Your critical insights and feedback have been invaluable to me. I hope we remain lifelong colleagues and friends. Srinjita and Varsha, I’m excited you decided to join us. Your enthusiasm for research adds an extra spring to my step. Heather, thanks for being the most awesome post-doc ever. I can’t express how thankful I am for the late nights you spent reviewing paper drafts on my behalf. Colleagues at UCAR: Holly, John and Jonathan, thank you for your kindness, support and encouragement over the years. This dissertation is only possible because of the truly magnificent work you’ve done in developing and maintaining the CCS platform.

I would also like to thank my thesis committee for their guidance, kindness and support throughout the process. I appreciate every minute you’ve afforded me from your hectic work schedules.

Finally, I would like to thank my parents, Isaac and Rose, and my brothers, Ruona and Tega. I am deeply indebted for your unconditional love and support in all my endeavors.
## Contents

### Chapter

1 Introduction

1.1 My thesis ......................................................... 1
1.2 The value of online learning communities ............................ 1
1.3 Using social network analysis to understand online learning communities ...... 2
1.4 Research context and scope ........................................ 3
1.5 Research questions ................................................. 4
1.6 Research design .................................................... 6
1.7 Research contributions ............................................. 6

2 Related Work

2.1 Sociological network theory ....................................... 8
   2.1.1 Homophily and tie strength ................................. 10
   2.1.2 Triadic closures ............................................. 11
   2.1.3 Weak ties in online social networks ...................... 11
2.2 Recommendation systems ......................................... 12
   2.2.1 Overview of the recommendation problem ................ 13
   2.2.2 Collaborative filtering recommendation ................... 13
   2.2.3 Content based recommendation approaches ............... 15
   2.2.4 Hybrid recommendation systems ........................... 16
2.2.5 Network analysis and recommendations ........................ 17
2.3 Understanding the diffusion of information ........................ 18
  2.3.1 Heavy-tailed distributions and Power laws .................... 20
  2.3.2 Resource position and diffusion .............................. 25
  2.3.3 Resource quality and diffusion .............................. 26
  2.3.4 Information cascades ....................................... 27

3 Research Context

4 Weak ties as a lens for understanding resource usage behaviors of an online community of educators
  4.1 Purpose ....................................................... 37
  4.2 Data ......................................................... 37
  4.3 Methodology .................................................. 38
  4.4 Results ....................................................... 41
    4.4.1 Analysis 1: Educators who use community-contributed resources ................................. 41
    4.4.2 Analysis 2: Deduced social network of weak ties .............................. 42
    4.4.3 Analysis 3: Evolution of the deduced social network .............................. 45
  4.5 Discussion .................................................... 49
    4.5.1 Implications for school districts and educators ............................. 50
    4.5.2 Benefits of mixed methods research .................................. 51
  4.6 Limitations ..................................................... 51
  4.7 Conclusion ..................................................... 52

5 Using network analysis of clickstream data to improve resource recommendation systems
  5.1 Purpose ....................................................... 53
  5.2 Data ......................................................... 53
5.3 Methodology ................................................................. 54
  5.3.1 Collaborative filtering resource recommendation system ............... 54
  5.3.2 Content-based resource recommendation system ........................ 55
  5.3.3 Hybrid resource recommendation system ............................... 56
  5.3.4 Triadic closure augmentations of both the collaborative filtering and hybrid recommendation systems ............................... 57
  5.3.5 Evaluating recommendation systems .................................. 58
  5.3.6 Online experiments ................................................... 59
  5.3.7 Using offline experiments to evaluate the recommendation systems developed ............................................................... 60
5.4 Discussion ................................................................. 63
  5.4.1 Triadic closures and trustworthy recommendations ...................... 63
  5.4.2 Limitations ............................................................ 64
5.5 Conclusion ................................................................. 65

6 An information cascade model for understanding the usage distribution of resources 66
  6.1 Purpose ................................................................. 66
  6.2 Data .................................................................... 66
  6.3 Methodology ............................................................ 67
  6.4 The usage distribution of community-contributed resources ............... 67
    6.4.1 User behavior and the observed heavy-tailed distributions .......... 69
    6.4.2 Usage distribution of community-contributed resources as a power law 70
  6.5 Mechanisms behind the power law distribution of resources ............... 75
    6.5.1 Resource position ................................................ 75
    6.5.2 Resource quality ................................................ 76
    6.5.3 Social influence and resource usage ................................ 78
  6.6 Results ................................................................. 87
6.7 Discussion ................................................................. 88
6.8 Limitations ................................................................. 89
  6.8.1 Size of the datasets .............................................. 89
  6.8.2 Logged position of community-contributed resources .......... 89
  6.8.3 Social influence at the individual level ......................... 90
6.9 Conclusion ................................................................. 91

7 Final thoughts .......................................................... 92
  7.0.1 Research contributions ......................................... 94
  7.0.2 Potential future directions ...................................... 96
7.1 Acknowledgements ...................................................... 99

Bibliography ............................................................... 100

Appendix

A CCS pre-deployment survey questionnaire ................................ 110
  A.1 Years teaching ....................................................... 110
  A.2 Class size ............................................................. 111
  A.3 Level of comfort and use of technology when teaching .......... 111
  A.4 Propensity to use and share resources .......................... 112
  A.5 Perceived level of isolation ....................................... 113
  A.6 Class needs .......................................................... 113
### Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Data sources employed for each research question and context of use. ‘n’ indicates the number of users that used community-contributed resources in the data time frame.</td>
</tr>
<tr>
<td>4.1</td>
<td>Categorization of survey items. A full list of all questions that make up each survey category is available in Appendix A.</td>
</tr>
<tr>
<td>4.2</td>
<td>Candidate triad features</td>
</tr>
<tr>
<td>4.3</td>
<td>Traidic closure prediction performance using J48 Decision trees.</td>
</tr>
<tr>
<td>5.1</td>
<td>Description of datasets under analysis. The 2011-2012 data is used in building the triadic closure prediction model (see chapter 4), and the recommendation systems are built and evaluated on the 2012-2013 data.</td>
</tr>
<tr>
<td>5.2</td>
<td>Description of features of a community resource</td>
</tr>
<tr>
<td>5.3</td>
<td>Average Precision at N for all five recommendation systems evaluated.</td>
</tr>
<tr>
<td>6.1</td>
<td>Description of clickstream datasets under analysis. NA indicates that this statistic does not apply to the year under observation</td>
</tr>
<tr>
<td>6.2</td>
<td>Comparison of the skewness and kurtosis of the usage distribution for each academic year</td>
</tr>
</tbody>
</table>
6.3 Correlation tests between the last, mode and median click position of resources and their eventual usage. Statistically significant correlations are highlighted in bold text. 76

6.4 Mapping of quality indicators of Bethard et al. [17] to signals that can be inferred before using a resource. An 'X' indicates that a signal for this indicator is not present. 78

6.5 Signal weights 78

6.6 Results of information cascade simulation experiments 87
Figures

2.1 Comparison between a normal and long-tailed distribution. A normal distribution is illustrated in (a) and a heavy-tailed distribution is shown in (b). 21

2.2 An illustration of a power law determined from a logarithmic transformation of the population of cities. This figure was adapted from the book Nature’s Patterns: Exploring Her Tangled Web [62]. 22

3.1 Snapshot of the CCS showing community-contributed resources. The various labels highlight user annotations (tags, saves and ratings) and features of these resources. 34

4.1 Deduced social network showing three sub-communities and bridge nodes. Anonymized node labels represent the user’s school with the grade level in parenthesis. 42

4.2 Depiction of the four possible states of a triad in our network. 46

4.3 Comparison between triadic formations in random graphs versus the deduced network. 47

6.1 Usage distribution across for each academic year. 68

6.2 Usage distribution across all academic years. 70

6.3 A comparison between the total number of resources used and the number of resources used before they had been saved per academic year. 71
6.4 A log-log plot of the number of resources (y-axis) and number of users (x-axis) constructed using logarithmic binning (bins of exponentially increasing widths) shows up as a roughly straight line. This provides preliminary though inconclusive evidence of a power law distribution. .......................... 72

6.5 Comparisons of the complimentary cumulative distribution function (CCDF) of the empirical data, the power law, lognormal and exponential distribution fits to the data. .............................................................. 74

6.6 Correlation between saves and usage over time ........................................ 79

6.7 Trend lines showing the relationship between the number of unique users and saves on a resource .......................................................... 80
Chapter 1

Introduction

1.1 My thesis

Sociological network theory can be used to explore the phenomena of homophily and triadic closures in online learning communities wherein ties between members are not evident. Consequently, an understanding of the triadic closure process can be used to improve the recommendation of resources in these communities. Finally, social influence may play a significant role in the diffusion and popularity of resources within online learning communities.

1.2 The value of online learning communities

Online learning communities are becoming a main-stay of K-12 teacher professional development. Research consistently indicates that these communities can improve the instructional practices of educators [101] and produce increases in student learning [81, 87] by providing educators with access to learning resources and best practices shared by their peers. For example, an educator may contribute a lab exercise which makes a specific scientific concept very accessible to English Language Learners that another educator from a different school with a similar student demographic will find useful enough to incorporate into her instruction. Given the importance of online learning communities to teaching and learning, understanding their dynamics—and the factors that influence these dynamics—is an intriguing question with important implications for educator instruction, student learning
and agencies that support online learning communities.

This dissertation shows that sociological network theory can be used to understand the dynamics of online learning communities. Specifically, this dissertation argues that the phenomena of homophily (tendency of individuals to have social ties with others of similar traits) and triadic closures (tendency of new connections to develop between individuals sharing a common neighbor) can be understood through the resource sharing and usage behaviors of educators. I also demonstrate how an understanding of the triadic closure process can be used to improve the performance of traditional resource recommendation systems. Finally, I show that social influence may play a significant role in the diffusion and popularity of resources within online learning communities.

1.3 Using social network analysis to understand online learning communities

Social network analysis (SNA) techniques have been widely used to understand the dynamics of physical and digital communities [50, 42, 48]. These techniques, in general, can be operationalized on communities with a defined network structure i.e. communities where nodes (individuals) and edges (ties) between nodes are known. A tie could be a friend or follower relation where, for a set of users $A$ and $B$, $A$ is a friend of $B$ or $B$ follows $A$ and vice versa. However, in communities where ties between members are not evident, SNA techniques are not directly applicable [4]. By deducing ties between members where such ties are not evident, this dissertation builds upon existing research [4] that has begun to make communities without immediately identifiable social ties amenable to SNA techniques.

Networks wherein ties between members are deduced are called Deduced Social Networks (DSNs). DSNs are object-centric, i.e. ties between nodes are created based on individual interest in the same set of objects, such as pages and resources. In this research, a DSN is constructed by creating an undirected edge $^1$ between two educators (nodes) that use the same community-contributed resource. Community-contributed resources include lesson

---

$^1$ A undirected edge denotes a symmetric (bi-directional) relationship between the nodes it connects.
plans, videos, animations, reading guides, assessments and other types of teaching resources that educators have found useful in their instruction, and have shared with others in the online learning community. These resources embody both the instructional content and the pedagogy of educators who share them [114]. I show that the phenomena of homophily and triadic closures in an online learning community can be understood through the resource usage behavior of its members.

1.4 Research context and scope

The research outlined in this dissertation is based on the resource usage and sharing patterns of an online community of middle and high school Earth science educators. This community of educators used a web-based instructional planning tool called the Curriculum Customization Service (CCS) [123] that was deployed within a large urban school district. Explained in full detail in chapter 3, the CCS provides educators with access to digital versions of their class textbook, digital library resources, and an online community where they can access and share resources with other educators. Prior research indicates that shared resources can support educators in customizing their instruction to meet the needs of students in their classrooms [123, 115, 92]

The core of this dissertation is divided into three phases. The first phase extended prior work [4] on DSNs by exploring the applicability of sociological network theory to understanding the nature of the deduced ties in the network of educators that used the CCS. In this phase, I used sociological network theory as a lens for understanding the phenomena of homophily and the formation of triadic closures within this community over time.

The second phase of this dissertation explored the suitability of computational models developed in phase 1 for developing improved resource recommendations systems within the CCS. Specifically, I used a computational model for predicting the formation of triadic closures to improve a traditional collaborative filtering system for resource recommendation.

The third phase of this research focused on understanding the underlying mechanisms
behind the diffusion and popularity of community-contributed resources. As with other domains that command an audience, such as books and movies, community-contributed resources are characterized by an extreme imbalance in user interest: a few resources are widely used, while most resources attract very little or no attention. My research specifically investigated whether the visibility of a resource (its position on the list of resources it is displayed in), its quality, or the usage behaviors of members of the learning community would provide insight into the resource’s diffusion.

This introductory chapter ties together each phase of this dissertation. In it, I give an overview of the research questions, the research design, and the primary contributions of this work. Chapter 2 examines relevant related research in sociological network theory, recommender systems, and network models for studying information diffusion and cascades. Chapter 3 discusses the Curriculum Customization Service (CCS). Chapters 4, 5 and 6 detail each of the studies conducted as part of this dissertation research. Finally, I conclude with my final thoughts and potential future directions for the research outlined in this dissertation in chapter 7.

1.5 Research questions

This dissertation explored the following questions:

(1) What theories are useful for understanding and characterizing the nature of a deduced social network of educators?

This dissertation studied an online learning community in which ties between educators were not evident. To understand the interactions between members of this community, this dissertation built upon the idea of a DSN by connecting educators based on their usage of community-contributed resources. Accordingly, this dissertation shows that sociological network theory can be used to understand the phenomena of homophily and triadic closures in a DSN. I discovered that educators with high
edge weights in the DSN were more similar to each other in comparison to educators with low edge weights. The edge weight between two educators is determined by the number of community-contributed resources they have used in common. For example, a pair of educators that have used two community-contributed resources in common will have an edge weight of 2. Furthermore, I discovered that the formation of new connections via the triadic closure process can be understood and predicted. Under certain conditions, an edge $A - B$ is likely to exist between two unconnected educators $A$ and $B$ that share an edge with another educator, $C$.

(2) What implications does an understanding of the deduced social network between educators have for resource recommendation?

I explored the utility of a computational model for understanding the formation of triadic closures to improving the performance of resource recommendation systems. My results indicate that traditional collaborative-filtering and hybrid recommender systems that are augmented with a triadic closure prediction model can provide more accurate resource recommendations than a basic collaborative-filtering or hybrid recommendation system.

(3) What underlying mechanisms best explain the diffusion of resources in the community of educators?

As in many other domains, resources in the online learning community of educators experience a phenomena of unequal usage equity. A few resources receive the bulk of user attention while most others get little to no attention. This inequity in usage leads to a heavy-tailed distribution in the usage of resources. I show that social influence exerted by the usage behavior of educators may be responsible for the observed inequity in the usage distribution of resources. This finding gives premise to agencies that support online learning communities for seeking alternating ways of promoting high quality learning resources that are more resilient to social influence.
e.g. personalized email recommendations.

1.6 Research design

This dissertation is comprised of three studies, which respectively address each of the research questions posed in section 1.5.

The first study investigated the applicability of sociological network theory to understanding the nature of the deduced social network between educators. The results of this study have been published in the Advances on Social Network Analysis and Mining (ASONAM) 2015 conference proceedings.

The second study explored the extent to which computational models developed in the first study are suitable to the improvement of resource recommendation systems. The results of this study have been accepted with revisions by the journal of Social Network Analysis and Mining (SNAM).

The third and final study of this dissertation explored the underlying mechanisms behind the usage distribution of community-contributed resources. The results of this study have been submitted to the conference on Educational Data Mining (EDM).

The methods, findings and limitations of each of the aforementioned studies are detailed in Chapters 4, 5 and 6 respectively.

1.7 Research contributions

This dissertation makes intellectual contributions to the fields of social network analysis, recommendation systems and economic theories on information cascades. I demonstrate how social network analysis techniques, coupled with sociological network theory can be used to understand the dynamics of communities wherein ties between members are not evident. Specifically, I show that sociological network theory can be used to understand the phenomena of homophily and triadic closures in the online community of educators. Consequently, this dissertation contributes novel techniques for improving traditional educational recom-
mender systems using computational models for predicting triadic closures in the deduced social network.

Finally, this dissertation contributes computational models for understanding information diffusion processes that lead to an inequity in the usage distribution of community-contributed resources. In particular, I extend the classic information cascade model of Bikchandani, Hirshleifer and Welch (BHW) to scenarios where individuals select among multiple choices and when only an aggregate of the decisions of prior individuals is available to them when making their selections.

In addition to the aforementioned intellectual contributions, findings from this dissertation will be beneficial to educational agencies, such as school districts that are invested in online professional learning communities. By coupling field research with usage log data, this dissertation highlights the value of online learning communities to educators. This can help broaden a school district’s understanding of the needs of its educators within an online learning community. Furthermore, an understanding of the factors that impact the diffusion of resources in online learning communities can be used to promote the usage of high-quality but barely used resources.
Chapter 2

Related Work

The literature and related work that frame this dissertation draw from three interdisciplinary areas of research. First, sociological network theory provides a lens with which the nature of the deduced social network is analyzed. This analysis underpins further exploration of network phenomena in the community of educators that is being studied. Second, current advances in educational recommendation systems are explored and contrasted with a novel resource recommendation system inspired by the findings from study 1. Third, models from economic theory on information cascades and social learning guide the exploration of mechanisms behind the diffusion of resources in the community of educators. I show that social influence—investigated through the lens of information cascades—may be a primary driver of the usage of community-contributed resources.

2.1 Sociological network theory

In this section, I explore Granovetter’s [66] theory on the strength of weak ties. Consequently, I examine the applicability of Granovetter’s theory as a theoretical lens for understanding the deduced social network of educators. My goal is to understand if the deduced relationship between educators (based on usage of the same community-contributed resources) constitute weak ties.

The concept of weak ties was originally proposed by Mark Granovetter [66]. He defined the strength of a tie as “a (probably linear) combination of the amount of time, the emotional
intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie [66].” According to this definition, weak ties can be broadly described as relationships between acquaintances and strong ties as relationships between individuals that share a closer bond e.g. family members, friends, teammates etc. Granovetter proposed the concept of weak ties as a way of addressing deficiencies in sociological theory connecting micro-level interactions to macro-level patterns observed in a network. He argued that although statistical and qualitative studies provided good insights into macro-level phenomena in communities, they do not provide sufficient insights into micro level interactions that led to the observed macro-level patterns. As an example, consider two users A and B where A is an acquaintance (weak tie) of B and vice versa. A and B both have a collection of close friends (strong ties), most of whom are also friends with each other. In this scenario, a weak tie can serve as an important bridge between the cliques of A and B, thus facilitating the spread of ideas and even the galvanization of both cliques towards a common goal—such as a community rally for higher wages. Thus, while statistical and qualitative studies provide information on the size and spread of the rally, they do not provide insights into how it evolved. Specifically, they do not consider the interpersonal relationships between individuals from disparate groups such as the groups represented by A and B that led to the mobilization of both groups towards a common goal. Granovetter [66] hypothesizes that an interpersonal relationship between A and B is responsible for the unification of both groups—represented by A and B—towards a common goal. Furthermore, Granovetter [66] suggests that weak ties can be important sources of information in a network. In a subsequent study, he discovered that most people learn about new job opportunities through weak ties [64]. Other studies have also highlighted the impact of weak ties as important conduits for knowledge transfer in organizations and learning communities [75, 111, 102, 85].

In this research, I consider whether the deduced relationships between members of the community constitute weak ties. A deduced relationship exists when two community members view or access the same resource. If these deduced relationships do constitute weak
ties, then other theorized network properties should also be manifest, namely homophily and triadic closures. I investigate if and to what degree these two properties are supported in the data. Specifically, I examine the degree to which homophily is predicted by tie strength, and the degree to which the evolution of the network can be predicted by the formation of triadic closures.

2.1.1 Homophily and tie strength

Granovetter [66] posits that the stronger the tie between two individuals, the greater the level of homophily between them. Homophily is the idea that similarity breeds connection. In real world social structures, people tend to have relationships with other individuals of similar backgrounds such as gender, interests and beliefs [94]. The underlying mechanisms behind homophily are selection and social influence, and these processes work in-tandem in a social system. Selection is the idea that people tend to select friends who share similar characteristics/interests, and social influence is the notion that over time people tend to adapt their behaviors to be more like their friends. The impacts of selection and social influence on group cohesion in social systems have been studied extensively [50, 110].

In the context of this research, the strength of a tie is defined by the edge weight between the nodes (educators) it connects. For a set of nodes \((A, B)\), the edge weight between \(A\) and \(B\) is determined by the number of resources \(A\) and \(B\) have used in common. i.e., the strength of a tie between educators is defined by the number of common resources they have viewed or used. I explore Granovetter’s claim of the relationship between the strength of a tie and the similarity between the parties it connects to understand if any qualitative similarities exist between educators using the same resources. For example, qualitative similarities between educators might include the needs of students in their classes, their perceived isolation or their comfort using technology.
2.1.2 Triadic closures

According to Granovetter’s theory, triadic closures are a core mechanism driving the formation of weak ties in a social network. The triadic closure property states that if two people in a network have a friend in common with whom they each share a strong tie, there is an increased likelihood of a weak tie forming between these two people [50]. Reasons behind the formation of triadic closures include opportunity (B & C are more likely to meet if they have a common friend A), trust (the joint friendship B & C share with A provides a basis for trusting each other) and incentive (B & C not being friends with each other puts a latent strain on the relationship between A & B and A & C).

This research examines whether triadic closures in the social network can be explained by random processes. It then examines the degree to which tie strength and other network properties can predict the formation of triadic closures. Should the triadic closure property hold for ties in the deduced social network, it will provide further support for viewing these ties as weak ties.

2.1.3 Weak ties in online social networks

Researchers have applied Granovetter’s theory to understanding the dynamics of user relationships in online social networks. These approaches have studied link prediction [88], friend recommendations [74], viral advertising [120], and the recommendation of educational resources [5].

Huang et al. [74] illustrate how an understanding of the formation and evolution of triadic closures in directed \(^1\) social networks such as Weibo and Twitter can be used to improve user recommendations. User recommendations, popularly known as People You May Know (PYMK), are a classic problem in social networks where providing useful recommendations to users is essential to the growth of the network [43, 70]. Similar to Huang et al. [74], I

\(^1\) A directed graph is one in which edges connecting nodes have a distinct direction between them.
use network typology features in predicting the formation of triadic closures in the deduced network of educators.

In addition to improving user recommendations in social networks, the relevance of the theory of weak ties to product adoption has also been investigated. Research indicates that the strength of a tie (relationship) between users $A$ and $B$ has a significant impact on the cascade of product adoptions from $A$ to $B$ and vice versa [50]. Numerous computational approaches for understanding and predicting tie strength have been developed [57, 35, 104]. Similar to these prior approaches, this research combines social signals to compute tie strength: the total number of resources shared between two nodes (educators) denotes the strength of the tie.

In learning domains, Akbar et al. have investigated the use of deduced social networks among users of an educational digital library [5]. In this work, the deduced social network is computed by connecting people that have used the same library resource. I use Akbar et. al’s [5] technique to compute the deduced social network. This research significantly extends the work of Akbar et al. [5] by studying the degree to which these deduced connections constitute weak ties according to Granovetter’s theory.

2.2 Recommendation systems

Over the last two decades, recommender systems have become an active area of research in academia and industry [3]. The high interest in this area is driven by the abundance of practical applications for it: from helping users deal with information-overload to improving the product sales of online marketers [118]. This section covers a range of research on recommender systems for online learning environments. I introduce each of the three main recommendation architectures: content-based recommendation, collaborative filtering and hybrid recommendation, and review implementations of these approaches in online learning environments. Finally, I discuss social networking augmentations to standard recommendation approaches including an introduction of the triadic closure prediction model developed
in the first phase of this research.

2.2.1 Overview of the recommendation problem

Generally speaking, recommendation engines try to estimate a user’s rating for an unseen item. This rating could be explicit such as the likert-scale rating a user might give an item, or implicit such as whether a user would use/purchase an item or not. In a learning environment with educational resources, the recommendation problem can be introduced as follows: Let $U$ be the set of all users and $R$ be the set of all possible resources. Let $f$ be a utility function that measures the usefulness of a resource $r$ to a user $u$, i.e., $f : U \times R \rightarrow \mathbb{Z}$. For each user $u \in U$ we want to select a resource $r \in R$ that maximizes the user’s utility. More formally: $\forall u \in U, r_u = \arg\max_{r \in R} f(u, r)$

2.2.2 Collaborative filtering recommendation

Collaborative filtering recommendations work by recommending resources to a user based on the choices of other users who have similar behaviors to the current user [107]. Formally, a collaborative-filtering recommender attempts to predict the utility $f(u, r)$ of a resource $r$ to a user $u$ based on the ratings of that resource by other users that are similar to $u$. The utility of a resource $f(u, r)$ to a user $u$ is estimated based on the utilities $f(u_i, r)$ assigned to $r$ by all users $u_i \in U$ that are similar to $u$. For example, an online book store may have a collaborative filtering system that recommends books to a user based on the selection of other users with a similar purchase or rating history. On sites like Amazon, collaborative filtering recommendations usually appear under the title “Customers like you also purchased these items.” The collaborative filtering recommendation system computes similarity between users based on a distance metric such as the Cosine or Jaccard distance metric.

Collaborative filtering approaches for resource recommendation have been investigated in several learning environments [23]. Early implementation of collaborative filtering rec-
ommender systems in online learning platforms include work by Recker et al. [108] where a collaborative filtering system called Altered Vista was developed to recommend resources to teachers and students using an online educational platform of learning resources. The Altered Vista system not only provided users with recommendations of resources, it also connected users with other users who shared similar interests. These connections enabled similar users to connect for collaboration [108].

Unlike the work of Recker et al. [108] where a collaborative filtering recommendation system was built solely based on user ratings of resources, more complex collaborative filtering recommendation approaches have also been investigated in learning domains. Manouselis et al. [89] show how a multi-attribute collaborative filtering recommendation system developed for the European version of the online learning platform SchoolNet can be used to provide users with useful high quality learning resources. In a multi-attribute collaborative filtering system, items can have multiple attributes, the ratings of which are used to predict the utility of an item to a particular user. For example, a multi-attribute collaborative filtering system can have ratings for a movie’s actors, director and genre. Thus, in recommending new movies to users who enjoyed and rated highly a particular movie, the recommender system considers the combination of ratings assigned to all attributes of the movie—its actors, directors and genre—by other users. Other multi-attribute collaborative filtering systems developed for e-learning platforms include work by Bobadilla et al. [23] where a learner’s score in addition to the usage of learning resources is used in calculating the similarity between users. The aim behind this metric is to discover learners that have both similar interests (in terms of learning resources used) and equivalent competence levels (in terms of learning scores achieved). Thus, learners with interests in the same types of resources who have equivalent levels of competence will be deemed more similar to one another as compared to those who just share an interest in similar resources.

Outside online learning environments, several variations of collaborative-filtering recommendation approaches have been implemented. For example, in the early 2000s, Amazon
introduced an item-item collaborative filtering algorithm which discovered items that tended to be purchased together as against the problem of finding similar users in a traditional collaborative filtering system [86]. This approach is reported to be more scalable and accurate in comparison to other recommendation approaches for Amazon’s specific domain [86].

2.2.3 Content based recommendation approaches

In comparison to collaborative filtering recommendation approaches, content based approaches find items similar to the set of items rated/used by the current user. Formally, a content-based recommender tries to predict the utility $f(u, r)$ of a resource $r$ to a user $u$ based on the utilities $f(u, r_i)$ assigned by user $u$ to resources $r_i \in R$ that are similar to the resource $r$. A content-based recommendation approach focuses on the properties of items and similarity between items is determined by measuring the similarity of their properties [107]. For example, a content-based movie recommender system would try to find other movies that are similar to the set of movies that have been highly rated by a user in the past. This similarity is determined based on the properties of movies such as the movie director, actors, genre, language and length. Two movies that share the same genre and director will be deemed as being more similar in comparison to movies that share none of these properties.

As with collaborative filtering approaches, content-based recommendation approaches have been explored in numerous learning platforms. A key area in online learning platforms where the application of content-based recommendation systems is prevalent is in the recommendation of resources that are of the same context as the currently viewed resource or the history of resources viewed by the user. For example, it is more appropriate to recommend Biology resources instead of Math resources to a learner working through a Biology lesson.

Gahauth et al. [55] show how a content-based recommendation system augmented with a good learner score improved the learning outcomes of students by providing them with higher quality resources. They describe a good learner score as a peer-review rating
mechanism that provides users with resources that have been used by good learners. Good learners are described as students who have achieved a score above 80% in a resource’s post-test evaluation [55]. The good learners score acts as an additional filter to the content-based recommendation engine by providing users with resources that are not only similar to the currently viewed resource, but have been used and evaluated by other users.

Other approaches have investigated recommender systems that address not only the content needs of learners, but also detect and address gaps in their understanding of domain concepts. Okoye et al. developed an educational recommender system that supported self-directed learning in Earth Science domains by identifying misconceptions in student essays and recommending resources to address those misconceptions [103]. Through a controlled study, they show that the Customized Learning Service for Concept Knowledge (CLICK) is able to improve the learning outcomes of students by promoting deeper reflection on science concepts and essay revisions that are indicative of students’ changing scientific understandings [103].

2.2.4 Hybrid recommendation systems

Hybrid recommendation systems involve a combination of both content-based and collaborative filtering approaches. Most online marketplaces and learning platforms feature hybrid recommender systems [107]. Hybrid recommender systems help overcome many of the limitations of content-based and collaborative filtering methods [13, 39]. Hybrid recommendation systems have been developed in several ways including: combining the results of separate implementations of collaborative filtering and content-based methods, incorporating features of a content-based system in a collaborative filtering approach and vice versa, and developing a unified model incorporating both content-based and collaborative filtering approaches [3]. Hybrid recommendation platforms incorporating the aforementioned styles have been investigated in numerous online learning platforms. They include learning systems that recommend online activities based on learner interests, habits and knowledge
levels [24] and systems that take into account the device (learning environment) from which recommended resources are consumed in making appropriate recommendations [11].

2.2.5 Network analysis and recommendations

Researchers have recently begun to investigate whether an understanding of the social connections between individuals in a network can be used to improve resource/item recommendations [12, 24]. Recommendation systems based on the social connections between individuals have been explored in both explicit social networks, i.e., where direct ties exist between users of the networks e.g. Facebook, Twitter, Academia.edu, and in implicit networks, i.e., networks where ties between users are deduced [36, 5]. Implicit connections are typically deduced from some kind of measurable user activity such as users clicking on the same resource, or replying to each other in a discussion forum.

In this research, I introduce an approach for performing network analysis on a social network of educators that has been deduced from the clickstream of their use of shared educational resources. In particular, I introduce a model for understanding the occurrence of triadic closures in the network and use this model to augment a traditional collaborative filtering recommendation approach by inferring future similarity between users. As described in section 2.1.2, triadic closures are a basic group phenomena in social networks that explain how new new edges (connections) are formed in a network. Triadic closures are one of the most basic principles for understanding the evolution of a network. The triadic closure property states that if two people in a network have a friend in common with whom they share a strong tie, there is an increased likelihood of the two becoming friends themselves [12]. In the context of this research, this can be interpreted as follows: if two educators A and B who haven’t used any of the same resources (thus do not share an edge), but have a common neighbor C with whom they share an edge (used a subset of the same resources), then there is an increased likelihood that A and B will be interested in the same set of resources in the future.
The principle of triadic closures has been successful used to improve user recommendations in directed social networks such as Twitter and Weibo [74, 26, 61]. A directed social network is one in which there is a direction to the relationship between any two members (nodes). For example, for a set of nodes $A$ and $B$ in a directed network, the connection between $A$ and $B$ is $A \rightarrow B$ if node $A$ follows $B$ and $B \rightarrow A$ if $B$ follows $A$. Triadic closures have also been used in predicting consumer choice in e-commerce platforms [67].

The principle of triadic closures has also been investigated in undirected social networks. In 2011, the popular streaming music service Spotify used the principle of triadic closures to improve its music recommendations to users based on common friends with whom they shared similar tastes [121].

This research explores the impact of a triadic closure prediction model on the recommendation of resources to an online community of educators using and sharing educational resources. To the best of my knowledge, this has not been investigated in any prior research.

2.3 Understanding the diffusion of information

Popularity is a convoluted yet striking network phenomenon characterized by extreme imbalances. For example, while most musical artists are known by a small fanbase, a few artists such as The Beatles, Michael Jackson, Coldplay and Madonna have attained global recognition. According to a 2013 report by the NextBigSound\(^2\), a whopping 90.7% of musical artists are undiscovered and only a meager 0.2% have attained mega status. A similar phenomena is observed in other domains such as books and movies[116, 50]. While a few books and movies have attracted widespread appeal, most books and movies are relatively unknown. In fact, we can personally relate to this phenomenon ourselves. Most of us go through life known only by a close circle of friends and family. However, a few of us attain widespread recognition e.g. Hollywood celebrities, politicians and royalty. As briefly

\(^2\) Launched in 2009 (now acquired by Pandora), the Next Big Sound is the leading provider of online music analytics and insights. It tracks hundreds of thousands of music artists and their fanbase around the world.
discussed in Chapter 1, a similar phenomenon is observed in the popularity distribution of resources in the community of educators. While a few resources are widely used (popular), most resources receive very little attention.

The reason behind this wide disparity in popularity has intrigued researchers for many years [116, 50]. Consider the wildly successful Harry Potter book series. If we could roll back time 20 years (before the release of the first novel) and then run history forward again would the Harry Potter series still be the juggernaut in children’s fiction that it is today, or would it languish in obscurity, allowing some other book to attain global precedence? This thought-provoking question has split researchers into two schools of thought. The first school of thought believes that qualitative differences between items ultimately determine popularity. Thus, the dominance of the Harry Potter series will be attributed to its higher quality in comparison to the alternatives. The second school of thought believes that social influence exerted by the decisions of individuals in the network is primarily responsible for the wide disparity in popularity [50].

To investigate both schools of thought on the reason behind the wide disparity in popularity among similar items, Salganik et al. [116] performed groundbreaking experiments on an artificial music market place of relatively unknown songs. They created a music download site with 48 relatively unknown songs where visitors to the site were given the opportunity to listen to a song and download it [116]. The twist in their study was that visitors to the site were assigned to one of nine parallel copies or “worlds” of the site. Each “world” of the site fell under one of two conditions: a social influence condition where visitors could see the aggregate download count (market share) of each song and an independent condition where no aggregate download count information was available. Eight of these worlds fell under the social influence condition and the ninth world fell under the independent condition. This world served as a control in their experiments. It provided insights into a song’s popularity in the absence of social information. The results of their experiments showed that the market share distribution of songs in the independent condition was much
more uniform in comparison to each of the other eight worlds under the social influence condition, which were all characterized by heavy-tailed distributions. Even more intriguing, the most popular songs were different in each world under the social influence condition. Using the market share of songs in the independent condition as a benchmark for quality, Salganik et al. [116] discovered a very weak correlation between quality and popularity in each of the worlds under the social influence condition. The best songs rarely did poorly and the worst songs rarely did extremely well. However, for songs of a specific market share (quality) in the independent condition, their popularity varied widely across all eight worlds under the independent condition. Salganik et al. [116] conclude by stating that social influence exerts an important albeit “counterintuitive effect on the formation of cultural marketplaces, the effects of which is reminiscent of information cascades” [116]. The more information participants had regarding the decisions of others, the greater agreement they seemed to display in their musical preferences [116].

The analysis performed in this dissertation follows the same principle as the investigations of Salganik et al. [116]. This dissertation aims to discover what generative mechanisms underly the usage distribution of resources in the online community of educators. It investigates whether resource quality [17], visibility [82] or social influence—investigated through the lens of information cascades [18, 10, 44]—ultimately explain this distribution.

Before delving into prior work exploring each of these hypotheses, I shed light on heavy-tailed distributions. I begin by exploring prior work on characterizing heavy-tailed distributions, and then I explore how this characterization can provide insights into how they are generated.

### 2.3.1 Heavy-tailed distributions and Power laws

In comparison to the well-known normal distribution (also known as Gaussian distribution or Bell curve), heavy-tailed distributions contain a significant probability mass in the tail of the distribution. Consider the sample probability distributions in Figure 2.1. Rare,
unusual events (e.g. events $\geq 3$ standard deviations from the mean) are more likely in a heavy-tailed distribution (Figure 2.1(b)) in comparison to a normal distribution (Figure 2.1(a)).

The central limit theorem[73] provides a natural mechanism for the generation of normal distributions. It states that the limit of the sum (or average) of a sequence of independent random quantities will be distributed normally [73, 50]. For example, if one measures a fixed physical quantity such as body weight, the variations across measurements will be approximately normal. While the central limit theorem provides a natural mechanism for the generation of normal distributions, a similar universal generative mechanism does not exist for heavy-tailed distributions.

However, a specific class of heavy-tailed distributions known as power laws are thought to be the result of feedback effects that arise from correlated decisions across a population [100, 109, 38]. Also known as Zipf, Pareto-Levy distributions, scale-free or scale-invariant [16] distributions, a quantity is said to follow a power law when the probability of measuring a particular value of that quantity varies inversely as a fixed power of that value [100, 38].

Figure 2.1: Comparison between a normal and long-tailed distribution. A normal distribution is illustrated in (a) and a heavy-tailed distribution is shown in (b)
quantity $x$ obeys a power law if it is drawn from a probability distribution $p(x) \propto x^{-\alpha}$ where $\alpha$ is known as the exponent or scaling parameter. In practice, a distribution rarely follows a power law for its entire range of values [100, 38]. Power laws are typically defined for the values of $x$ in a distribution for which $x_i \geq x_{\text{min}}$ where $i = 1...n$ are the observed values of $x$ in the distribution [38]. $x_{\text{min}}$ is referred to as the lower bound of the power law distribution.

Power laws are rife in a wide array of man made and natural phenomena [100]. For example, the fraction of telephone numbers that receive $k$ calls per day is roughly proportional to $1/k^2$; the fraction of books that are bought by $k$ people is roughly proportional to $1/k^3$; the fraction of scientific papers that receive $k$ citations $1/k^3$; the frequency of use of words in English is roughly proportional to $1/k^2$, and the population of U.S. cities is roughly proportional to $1/k^2$ [100].

![Figure 2.2: An illustration of a power law determined from a logarithmic transformation of the population of cities. This figure was adapted from the book Nature’s Patterns: Exploring Her Tangled Web [62]](image)

Traditionally, power laws have been determined from a logarithmic transformation of the values taken by a quantity and their frequencies (see Figure 2.2). However, recent
research indicates that this approach to determining power laws is not fail-safe [100, 38].
In particular, a similar logarithmic transformation will appear as a straight-line for other probability distributions such as the log-normal\(^3\) and exponential distribution [100, 38].

Clauset et al. [38] provide a principled statistical framework for determining whether an empirical dataset follows a power law distribution. Their approach uses maximum likelihood fitting methods to determine the lower bound \(x_{min}\) and the scaling parameter \(\alpha\) for which a distribution follows a power law [38]. Consequently, they use goodness-of-fit tests to determine whether alternative probability distributions such as the exponential and log-normal distributions provide a better fit to the data. The approach of Clauset et al. [38] is covered in greater detail in the third study of this dissertation (section 6.4.2). I employ their approach to determine if the usage distribution of community-contributed resources follows a power law distribution.

2.3.1.1  Power laws and Rich-Get-Richer dynamics

Research indicates that entities that follow a power law distribution experience what is known as the “rich-get-richer” or “preferential attachment dynamic.” The general idea behind the rich-get-richer dynamic is that popular entities tend to get even more popular over time. Future popularity is directly proportional to current popularity. For example, individuals with a large following, such as celebrities, tend to garner an even larger following over time. In the context of this research, a community-contributed resource with an early lead in market share (number of unique users) over others will tend to extend this lead over time. The term preferential-attachment is commonly used in the literature on networks to describe the network growth process where new vertices (nodes) preferentially attach to vertices that are already well-connected i.e. vertices that have large number of other vertices connected to them.

To further explain the process of the rich-get-richer dynamic, I briefly describe the

\(^3\) A log-normally distributed quantity is one whose logarithm is normally distributed
“copying model” [50, 16, 97]—a simple generative mechanism behind the creation of web pages that produces the observed power law distribution of the in-degree\(^4\) (popularity) of web pages. As mentioned in section 2.3.1, the in-degree of web pages follows a power law of approximately \(1/k^3\).

**Copying model for rich-get-richer dynamics on the web**

1. Suppose webpages are created in a sequential order 1,2,3,..., N

2. When a page \(i\) is created, it creates a link to an earlier web page with a probability \(p\) where \(0 < p \leq 1\). This link creation can occur in one of two ways:

   - With a probability \(p\), page \(i\) creates a link uniformly at random to a page \(j\) from the set of all earlier created pages
   - With a probability \(1 - p\) page \(i\) creates a link to a page \(j\) where the probability of choosing page \(j\) is directly proportional to \(j\)’s current number of in-links. In this scenario the linking process is more likely to copy the decisions of earlier pages. Another way to understand this is to consider the prior step in which page \(i\) created a link uniformly at random to a page \(j\). Instead of creating a link to the page \(j\) chosen at random, the “copying model” assumes that page \(i\) creates a link to the page that \(j\) links to. Thus, copying the decision of \(j\).

Running this copying model for many pages has been shown to lead to a power-law distribution for the in-links of pages. Outside the web, the copying model has been used to explain the number of copies of genes, which has been shown to follow a power law distribution [78, 100, 31]. Since gene duplication arises from random mutations of DNA [1], then a gene which already has many copies is proportionally more likely to be lying in a random stretch of DNA that gets duplicated [50]. Thus, rich genes (those with many copies) tend to get even richer.

\(^4\) Number of pages that point to a current page
The rich-get-richer model only gives us an idea of how power laws evolve over time. It is based on the observable consequences of the decision making process [50]. This research goes a step beyond the rich-get-richer model to understand how power laws can arise from the decision making process of individuals. Kleinberg and Easley describe this as an “open and very interesting research question” [50].

I now address each of the aforementioned mechanisms—resource position, resource quality and models of information cascade—that may underly the usage distribution of resources in the community of educators.

2.3.2 Resource position and diffusion

Recent research indicates that the diffusion of a resource may be related to its rank or position as it is presented to a user. A recent paper by Lerman and Hogg on the social news aggregation site Digg indicates a strong correlation between the presentation order of a news story and its diffusion in the Digg network [83]. This correlation persisted irrespective of the aggregate social influence of the story. They reported that Digg users were more likely to engage with content that was highly visible (positioned at the top of the list) regardless of what signals of the content’s utility to others could be observed. This finding is in stark contrast to the work of Salganik et al. [116] which showed that the order of a resource had no influential bearing on its diffusion. The findings of Salganik et al. [116] indicate that aggregate social influence factors as described in section 2.3 have a more significant impact on the attention received by a resource. However, prior work on user behavior on web search engines indicate that the ranking of search results has an impact on user attention [27]. Items with a high ranking receive disproportionately more attention compared to lower ranked items [27].

In this dissertation, I examine the impact of a resource’s position on its diffusion in the community of educators. In particular, I investigate if the most frequently used resources (i.e. resources at the tail of the distribution in Figure 2.1(b)) are more visible in comparison
to other resources.

2.3.3 Resource quality and diffusion

Characterizing the quality of an educational resource is a non-trivial task for both humans and machines [17]. Bethard et al. [17] conducted a comprehensive study of an educational digital library to identify key indicators of resource quality that are both agreeable to humans and amenable to automation. The outcome of their study was a set of seven indicators that captured the rich, multi-faceted nature of quality in educational resources [17]. These indicators were whether the resource: had a sponsor, had a prestigious sponsor, had instructions, identified learning goals, identified age range, is appropriate for age range and is organized for learning goals [17]. They cover a broad spectrum of what educators attend to in evaluating a resource’s quality.

While Bethard et al. [17] examined quality based on the full content of a resource, I am interested in characterizing quality based on signals that can be inferred by users before deciding to use (click on) the resource. Specifically, I am interested in how a resource’s metadata—visible in its listing (see Figure 3.1)—give insights into its quality. I map quality signals inferable from this metadata to the indicators of Bethard et al. [17] to determine resource quality. I then examine the impact of a resource’s quality on its usage.

2.3.3.1 Why popularity is commonly seen as an indicator of quality

It is common for popularity to be viewed as a direct reflection of quality. This intuition is generally based on the concept of “wisdom of the crowds.” Wisdom of the crowds is the idea that an aggregate of the independent decisions of members of a group is strikingly accurate [124]. Empirical research based on this idea dates back to Francis Galton’s *Vox populi (The wisdom of the crowds)* [54]. Galton analyzed the responses of 800 participants at a weight-judging competition where participants attempted to guess the weight of an ox. He discovered that the guesses of most participants were either too high or too low. However,
the average of all guesses which he referred to as the *Vox Populi* (voice of the people) was surprisingly accurate—just 0.8% off from the actual weight of the ox. More recent studies have confirmed this phenomenon. After analyzing data from the popular quiz T.V. show *Who Wants to Be a Millionaire*, on which contestants can consult the studio audience or place a call to a trusted friend for help on a question, Surowiecki discovered that the studio audience performed much better in helping participants pick the right choice [124]. He found that the audience picked the right answers at a remarkable 91% accuracy compared to 65% of the time when participants consulted a friend [124].

Wisdom of the crowds is based on the linchpin of independence in the decision making process. In the context of this research, educators are able to observe the decisions of those before them if those educators saved the resources they used. Thus, an educator’s decision to use a community-contributed resource cannot be thought of as being independent.

### 2.3.4 Information cascades

Also known as observational or social learning, information cascades are defined as conformity that ignores one’s own tastes, beliefs or intuition [18]. They describe situations where it is optimal for individuals to ignore their personal preferences and imitate the decisions of those before them. For example, suppose you have a favorite restaurant *A*. While visiting a new town you discover that no one is eating at *A* but restaurant *B* next door is almost full. If you believe the other diners have tastes similar to yours, you may decide to ignore your preferred restaurant *A* to dine at *B*. In this scenario, the aggregate decision of others to dine at restaurant *B* has influenced you to also dine at *B*. This behavior is typically referred to as “herding” or “following the crowd.” Information cascades occur when people make decisions sequentially, with the latter individuals able to infer/observe the choices made by others before them [50].
2.3.4.1 Information cascades and the natural inclination to imitate

Prior research indicates that information cascades are intrinsically linked to animal and human nature to imitate others. For example, female guppies are most likely to choose male mates that they have observed being selected by previous females [56]. In humans, multiple studies have shown that infants early on in life, learn by mimicking the facial expressions and behaviors of adults [95, 7, 76].

Even in situations where it is more beneficial to differentiate, research indicates that individuals are still likely to imitate the decisions of others [77]. Kennedy [77] examined the business decisions of TV networks to introduce new programming from 1960-1989. Common business logic suggests that the introduction of a medical drama by ABC, for example, should reduce the benefit to NBC and CBS from doing so. However, if NBC and CBS believe that ABC has information about changing public tastes for different kinds of shows, they may want to imitate ABC’s choice and also introduce TV shows of the medical dram genre. After controlling for other factors, Kennedy found that TV networks tend to make introductions in the same categories as their rivals. He argued that this behavior is surprising given that businesses traditionally follow a “differentiation hypothesis” for success rather than the strategic imitation observed [77, 18]. Furthermore, other businesses such as banks tend to follow the same imitation strategy. After using census tract-level socioeconomic data, land-use data, and crime statistics to control for a census tract’s expected profitability, Chang and Jayaratne [32] show that a bank’s decision to open a branch in a census tract within the state of New York depended on the number of existing branches in that tract. The more branches a census tract had, the higher the likelihood that new branches would be opened in that tract. On a more disturbing note, research indicates that individuals are more likely to imitate decadent behaviors if they perceive the likelihood of punishment to be low relative to potential gain [32]. Individuals are more likely to commit crimes when they see others get away with it. Kahan [32] suggests that crime prevention and deterrence policies should take
social influence into account. He suggests that policies such as curfews and anti-loitering laws that make criminal activity less visible—thus reducing its social influence—should be implemented by law enforcement.

Information cascades can also be triggered intentionally. In 1995, management gurus Michael Treacy and Fred Wiersema propelled their book, *The Discipline of Market Leaders*, to the New York Times best seller list despite having mediocre reviews [21]. Treacy and Wiersema secretly purchased 50000 copies of their book from book stores whose sales were monitored by the New York times. Consequently their book made the New York bestseller list and sold well enough to continue as a best seller without further intervention from the authors [18].

### 2.3.4.2 Fragility of information cascades

Cascades can be as easily derailed as they can be started [18, 50]. Once an information cascade has begun public information stops aggregating i.e., the preferences of individuals in a cascade cannot be inferred from their decisions. As an example, consider the scenario where diners chose between two restaurants $A$ and $B$ in section 2.3.4. When individuals with a preference for restaurant $A$ begin going to $B$ simply because it is more popular, then their decisions stop contributing to the pool of knowledge regarding restaurant $B$. However, the cascade of diners going to $B$ can be easily broken. Suppose a revered restaurant critic attends both restaurants $B$ and $A$, and afterwards writes a poor review on restaurant $B$. Subsequent diners who read the restaurant critic’s review may decide to forgo $B$, despite its popularity, and attend restaurant $A$ instead. This may lead others to follow suit and stop patronizing $B$, and thus ending the cascade of goers to $B$.

### 2.3.4.3 Classical and modern studies of information cascades

Classical studies on information cascades include experiments by Milgram et al. [96], where they had groups of people stand on a street corner and stare up into the sky, and they
observed how many passersby stopped and also looked up at the sky. They found that with only one person looking up, very few passersby stopped. When the number of participants looking up was increased to five people, more passerby stopped, and when all participants looked up, they found 40% of all passersby also stopped and stared up into the sky. To understand this experiment through the lens of information cascades, we can assume most passerbys’ personal preference was not to look up at the sky. However, as more people stood up to look at the sky, many passerbys decided to override their personal preference and look up at the sky as well.

Modern studies on information cascades are based on the model of Bikchandani, Hirshleifer, and Welch (BHW) [18]. The BHW model uses Bayes’ theorem to show how individuals making binary decisions could end up in a cascade under rational conditions. The BHW model is made up of three core components: private signals, public signals, and a sequential decision making process. A private signal represents an individual’s personal preference. Private signals can be High (H) or Low (L). If an individual observes a High private signal for an action, then with a probability \( p \), he follows this signal and with a probability \( 1 - p \), he does not follow this signal. The converse occurs when the individual observes a Low private signal. Public signals represent the observable decisions of individuals. Individuals can make inferences on the private signals of those before them based on the public signals they observe. I illustrate the cascade process by paraphrasing an example from Bikchandani et al. [18]. Suppose three individuals: James, Kelly and Henry are sequentially deciding whether to adopt a new innovation. They each have a private signal (personal preference) as to whether adoption is a good idea or not. If they observe a High private signal then with a probability \( p > 1/2 \) they adopt the innovation. Individuals can make inferences on the signals of those before them based on their decision to adopt or reject. James starts off the process. He observes a High private signal and proceeds to adopt. Next in line, Kelly infers that James had a High private signal given his decision to adopt. Her decision to adopt is based on her private signal and the signal she inferred from James’ decision. If she
also observed a High private signal, she follows James’ decision to adopt. If she observed a Low private signal, she flips a coin and decides. Things get interesting with Henry. Suppose he observed that both James and Kelly had High private signals from their decisions to adopt. If he observes a High private signal, he follows suit and adopts as well. However, if he observes a Low private signal, he essentially computes the likelihood of adopting given three signals: \( \Pr(\text{adopting}|H, H, L) \). If the resulting value is \( > 1/2 \), he ignores his Low private signal and adopts like the others. This heralds the start of an information cascade. In controlled laboratory experiments, Anderson and Holt [10] confirm the accuracy of the predictions of the BHW model. 84% of the time, when individuals faced a situation like Henry’s, they proceeded to ignore their private signals and follow the crowd.

I extend the BHW information cascade model in three ways. First, my model accounts for cascades where individuals have more than two choices. In the CCS, educators have a choice between multiple community-contributed resources, and thus a cascade model based on binary choices is not sufficient.

Second, my model allows for cases where an individual decides not to use any of the alternatives. In the BHW model, individuals decide between two choices. However in many situations, there is a possibility that none of the available choices will be chosen. For example, an educator may decide not to use any of the available community-contributed resources.

Finally, my model incorporates cases where a public signal is not always available after a decision has been made. In the example of James Kelly and Henry making decisions in sequence, a public signal is observable from their decision to adopt or reject. Consequently, subsequent individuals can infer their private signal based on this public signal. In the context of this research however, the decision of an educator to use (click on) a resource does not leave a public trace. The only way subsequent educators are able to make inferences on the decisions of others before them is if these users contributed social metadata to a resource by rating or saving it. Only a summary statistic of the actions of predecessors is available to latter educators. Although their model works with a sequence of all actions
of prior educators, Bikchandani et al. [18] suggest that summary information can also trigger information cascades. Consider the automobile industry. An individual interested in a traditional sedan may learn that Toyota Corollas are outselling Honda Accords without knowing the order in which the vehicles of both brands are purchased. Imagine they watched a Toyota ad stating that 15 million Corollas had been sold compared to 10 million Honda Accords. Bikchandani et al. [18] suggest that the observability of summary statistics still leads to idiosyncratic outcomes, fragility, and cascades. The basic intuition is as before. Information keeps accumulating until a preponderance of evidence supports one action or the other by just enough to outweigh one individual’s private signal. At this point a cascade starts and new public information stops accumulating.

This research will examine if the summary statistics of the number of users that have saved a resource in the past can trigger information cascades as predicted by the BHW model [18].

2.3.4.4 Information cascades and power law distributions

In addition to the work of Bikchandani et al. [18], other works have investigated the distribution of choices that arise as a result of information cascades [45, 46, 44]. Devany and Walls discovered that the distribution of box office revenues of Hollywood movies during the first few weeks of opening follows a Pareto-Levy (Power law) distribution [45]. They hypothesize that information cascades may be responsible for hits—very successful movies—and bombs—movies that fail to impress—in the box office [45]. The reasoning behind this hypothesis is fairly straightforward. When people see a movie they like, they tell their friends about it, and encourage them to see it as well. Likewise, when they view a movie they do not like, they discourage their friends from seeing it. Over time, as audiences sequentially flock to the theaters, the information transmitted based on the experiences of others can lead to an information cascade.

Devany and Lee [46] developed an information cascade model simulating how audiences
choose between movies to see. Their goal was to investigate if this model would lead to a power law distribution of choices as hypothesized by Devany and Walls [45]. Devany and Lee [46] extended the BHW model to scenarios where agents choose between multiple movies and are able to observe both the decisions of local agents (agents before them), as well as an aggregate statistic of the market shares of all movies. The aggregate market share for all movies is global information which individuals may learn from reports on box office earnings in the news. As predicted by Devany and Walls [45], the model of Devany and Lee [46] did result in a Pareto-Levy distribution of choices.

In this dissertation, I show that an information cascade model can lead to a power law distribution in the usage of community-contributed resources. However, unlike the work of Devany and Lee [46], agents in my model only have access to a summary statistic of the decisions of others. In the context of this research, this summary statistic is the number of prior users that have saved a resource. I show that even in the absence of local information (the decision of agents before them)—as in both the models of Bikchandani et al. [18] and Devany and Lee [46]—information cascades can occur.
Chapter 3

Research Context

Figure 3.1: Snapshot of the CCS showing community-contributed resources. The various labels highlight user annotations (tags, saves and ratings) and features of these resources.

This dissertation research is based on educators’ use of an online instructional planning platform called the Curriculum Customization Service (CCS). The CCS (see Figure 3.1) provides middle and high school educators with access to digital versions of their class textbook, digital library resources and community-contributed resources [123]. Educators can organize these digital resources into personal collections and they can contribute resources for use by
others. They can also enhance resources by adding annotations such as tags and ratings. This dissertation is focused on educators who shared and used community-contributed resources. Community-contributed resources include lesson plans, videos, animations, reading guides, assessments and other types of teaching and learning resources that educators have found useful in their instruction.

The analyses conducted were based on the clickstream data and qualitative evaluations (interviews & surveys) of educators from a large urban school district that used the CCS. Clickstream data are the click actions of educators in the CCS—mouse clicks—which were anonymously tracked as a way to analyze their behaviors. This dissertation is based on the clickstream of educators over four academic years: 2009-2010, 2010-2011, 2011-2012 and 2012-2013. I now describe how both the clickstream and qualitative evaluations of educators were used to conduct the studies outlined in this dissertation.

The first study of this dissertation focused on applying sociological network theory to understanding the deduced social network and is based on two core datasets: clickstream data from the 2011-2012 academic year and a pre-deployment survey responded to by 37 educators prior to the 2011-2012 academic year. This survey asked a series of questions about teacher classroom settings, instructional practices, beliefs and needs. The survey instrument was developed by education and evaluation researchers and was validated prior to its use [80, 129, 28]. I used the results of this survey to investigate if homophily between educators could be determined based on their connections in the deduced social network. Specifically, I investigated whether higher edge weights between connected nodes signified a stronger presence of homophily.

The second study used clickstream data from the 2011-2012 and 2012-2013 academic years. A computational model for predicting the formation of triadic closures in the network was based on the 2011-2012 school year. The applicability of this model to improving traditional recommendation systems for resource recommendation was investigated on the 2012-2013 dataset.
The final study of this dissertation used clickstream data from all four academic years. The usage distribution of community-contributed resources was investigated for each academic year, and longitudinally over all four years. Table 3.1 summarizes the datasets and methods used in this dissertation.

Table 3.1: Data sources employed for each research question and context of use. ‘n’ indicates the number of users that used community-contributed resources in the data time frame.

<table>
<thead>
<tr>
<th>Research questions</th>
<th>Data used</th>
<th>Methods to be employed</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ 1</td>
<td>Clickstream data from the 2011-2012 academic year (n = 42)</td>
<td>Clickstream data was used to construct the deduced social network. An edge was created between two educators that clicked on the same community-contributed resource.</td>
</tr>
<tr>
<td></td>
<td>Pre-deployment survey data from 2011-2012 academic year (n = 37)</td>
<td>T-tests were used to identify similarities and differences between groups of educators based on survey items. These groups were determined by categorizing educators based on the edge weights they shared in the deduced social network.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Machine learning classifiers were developed to predict the formation of triadic closures in the deduced social network.</td>
</tr>
<tr>
<td>RQ 2</td>
<td>Clickstream data from the 2012-2013 academic year (n = 55)</td>
<td>The triadic closure prediction model from RQ 1 was used to develop a recommender system that was tested and evaluated on the 2012-2013 clickstream data.</td>
</tr>
<tr>
<td>RQ 3</td>
<td>Clickstream data from all school years (2009 - 2013) (n = 154)</td>
<td>The usage distribution of community-contributed resources was investigated for each academic year and over all four academic years.</td>
</tr>
</tbody>
</table>
Chapter 4

Weak ties as a lens for understanding resource usage behaviors of an online community of educators

4.1 Purpose

The purpose of this study is to understand the nature of the deduced social network created from the community-contributed resource usage behaviors of educators in the CCS. As a reminder, the deduced social network is generated by creating an edge between two educators that click on the same community-contributed resource. I examined the degree to which the phenomena of homophily and triadic closures as predicted by sociological network theory are exhibited by the deduced social network.

The findings of this study will provide insights into the resource usage behavior of educators and how participating in the online learning community supports the instruction of educators.

4.2 Data

This study was carried out using a mixed-methods approach that combined both qualitative data (survey) and quantitative data (clickstream of educators). The analyses performed are based on two distinct datasets. They are:

- The clickstream data of educators (n=42) who used community-contributed resources in the CCS during the 2011-2012 school year.
• A pre-deployment survey completed by 37 educators who used the CCS during the 2011-2012 academic year. Responses from 37 of the 130 items in the pre-deployment survey were grouped into six categories that provide insights into the instructional practices, classroom needs and beliefs of educators. The remaining survey items asked questions about CCS usage, and are therefore not relevant to the research being performed here. The six categories covering the instructional practices, classroom needs and beliefs of educators are: class size, years teaching, class needs, level of comfort and use of technology, perceived level of isolation from other Earth Science educators, and propensity to use and share resources of educators. These categories, along with a description of the survey items and resulting data, are shown in Table 4.1. Furthermore, the questions which each category is composed of are listed in Appendix A.

4.3 Methodology

A core goal of the CCS is to support the development and enhancement of a school district’s professional learning community. The primary way in which the CCS facilitates a school district’s professional learning community is by enabling educators within the district share teaching and learning resources that they have found useful or created with other teachers.

Determining the degree to which the CCS achieved these aims is one of the motivations for the research presented here. Towards this end, this study examines the concept of weak ties deduced from the community-contributed resource usage patterns of educators in the CCS’s online learning community. Specifically, study builds on Granovetter’s [66] theory of weak ties as described in section 2.1. As a reminder, weak ties are relationships between acquaintances; these relationships have a lower trust barrier and are easier to maintain as compared to strong ties, which are characteristic of the relationship between family, close friends and colleagues. Weak ties provide important pathways for new information in a net-
work [50, 66]. In contrast, redundant information is often circulated among those who share strong ties. This is primarily due to the frequency and intensity of the interactions between those connected by strong ties [66, 64]. For example, over time, two Earth science educators at the same school may develop a significant amount of shared background knowledge from frequently exchanging ideas and resources with one another. However a weak tie to one of these Earth science educators may expose both of them to new ideas and methods for improving their instruction.

I deduced the latent social network formed between educators as follows: I hypothesize a weak tie (undirected edge) to exist between two users (nodes) who use the same community-contributed resource. Such an edge indicates mutual interest in the same set of resources. If this hypothesis is valid, i.e., that deduced relationships constitute weak ties, then other properties of Granovetter’s theory should hold [66]. Specifically, I examine the degree to which homophily is predicted by tie strength, and the degree to which the evolution of the network can be predicted by the formation of triadic closures. I also explore whether any differences exist between educators who are part of the deduced social network and those who are not.

This study is addressed through the following research questions:

1. Are there qualitative differences between educators who use community-contributed resources and those who do not? Specifically, do differences exist in the beliefs, instructional practices and teaching needs of educators who use community contributed resources and those who do not use community-contributed resources?

2. What is the nature of the deduced social network between educators using community-contributed resources? What is the relationship between the strength of a deduced tie and the level of homophily between the educators it connects?

3. How does the network of educators who use community-contributed resources evolve over time? Does this network exhibit the triadic closure process indicated by Gra-
novetter’s [66] theory?

Table 4.1: Categorization of survey items. A full list of all questions that make up each survey category is available in Appendix A.

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class size</td>
<td>A numeric value indicating average class size</td>
</tr>
<tr>
<td>Years teaching</td>
<td>Respondents reported on their total number of years teaching in general and their total number of years teaching Earth Science in particular.</td>
</tr>
<tr>
<td>Class needs</td>
<td>A set of six Likert-Scale questions (responses ranging from strongly disagree to strongly agree) that asked educators about the presence of students with different reading abilities, quantitative skills, different cultural backgrounds and gifted and talented (GT) students in their classrooms.</td>
</tr>
<tr>
<td>Level of comfort and use of technology when teaching</td>
<td>A set of sixteen Likert-Scale questions that inquire about the use of various computer technologies in the classroom. These questions sought information on educator usage of social networking sites, digital libraries etc and their level of comfort with these technologies. Questions asked also included teacher frequency of use and comfort level with data processing programs such as Microsoft Word and Excel, social networking sites, search engines and streaming services.</td>
</tr>
<tr>
<td>Perceived level of isolation</td>
<td>A set of six Likert-scale questions that inquire on the opportunities of educators to interact with peers in their school district and their awareness of the practices of other Earth Science educators. Questions asked include teachers’ rating on the opportunity to interact with other Earth Science educators; attend relevant workshops and conferences; share materials with other educators; use materials of other educators; look at materials of other educators for inspiration and awareness of the instructional and curriculum planning practices of other Earth Science educators in their district.</td>
</tr>
<tr>
<td>Propensity to use and to share resources</td>
<td>A set of six Likert-Scale questions that inquire on the comfort level, frequency and ease with which educators shared teaching resources with peers in the same school district.</td>
</tr>
</tbody>
</table>

4.4 Results

A set of three analyses were performed to address each of the questions posed in section 4.3. The findings of each of these analyses are highlighted in this section.

4.4.1 Analysis 1: Educators who use community-contributed resources

This study addresses the first research question on the differences between educators that use community-contributed resources and those who do not. It investigates whether any differences exist in the teaching beliefs, needs and practices of these two groups of educators.

The set of teachers that participated in the survey were split into two groups: educators who used community-contributed resources (n=18) and educators who did not use community-contributed resources (n=19). Single-tailed independent t-tests were performed between both groups for all categories of survey items shown in Table 4.1.

The hypothesis for each t-test is as follows: “For a category $x$, educators who used community-contributed resources express a higher average rating compared to educators who did not use community-contributed resources.” The null hypothesis is that there is not a significant difference between the ratings of both groups. As an example, for the propensity to use and share resources category, I performed a t-test on the following research hypothesis: “Teachers who used community-contributed resources have a higher propensity to use and share resources compared to teachers who did not use community resources.”

T-tests as described in the preceding paragraph were performed for each category listed in Table 4.1. A statistically significant result was found for the following hypothesis: “Teachers who used community resources have a higher perceived level of isolation in comparison to teachers that do not use community resources.” ($t = 1.761, P = 0.0434$ at $p < 0.05$). No other statistically significant results were found between educators that used community-contributed resources and those that did not for all other category variables.
This finding is interesting as it supports a core premise behind the school district’s adoption of the CCS: to promote the sharing of knowledge and best practices within the professional learning community (PLC) of its educators. Research indicates that the lack of a PLC that promotes knowledge sharing can contribute to educators’ feeling isolated in the classroom. Results suggest that educators with higher levels of perceived isolation are more likely to use, and potentially benefit from online knowledge sharing with their peers. Consequently, this sharing may lead to improvements in teaching performance as highlighted by prior research on teacher learning [34, 99].

4.4.2 Analysis 2: Deduced social network of weak ties

![Figure 4.1: Deduced social network showing three sub-communities and bridge nodes. Anonymized node labels represent the user’s school with the grade level in parenthesis.](image)

This study addresses the second set of research questions that are concerned with understanding the nature of the deduced social network of educators. The social network
is deduced by creating a tie (undirected edge) between two educators (nodes) who used the same set of community-contributed resources. The full deduced social network at the end of the 2011-2012 school year is shown in Figure 4.1.

The analysis reported here was carried out in two steps. First, I determined the presence and characteristics of sub-communities in the network. Second, this analysis investigated whether there is a direct relationship between the strength of a deduced tie and the level of homophily between the nodes (educators) it connects.

I now describe each of these tasks and the analyses that were performed

### 4.4.2.1 Detecting sub-communities in the social network

The first task in understanding the nature of the social network was to discover the presence and characteristics of sub-communities in the network. This analysis was based on the clickstream data of educators who used community-contributed resources as explained in section 3. There were 580 unique click actions on 372 community-contributed resources. Irrespective of the number of times an educator clicks on a specific resource, I define this to be one unique click action.

Consequently, the Clauset-Newman-Moore’s [37] clustering algorithm was employed to discover community structures in the network. Clauset-Newman-Moore is a fast hierarchical clustering algorithm suitable for very large graphs [37]. Communities within the network are detected based on a property of graphs called modularity. Modularity is a measure of the likelihood of modules (communities) within a graph. A network of high modularity indicates dense connections between nodes in the same module and sparse connections between nodes of different modules. Running Clauset-Newman-Moore’s clustering algorithm on the deduced social network yielded three distinct sub-communities: G1, G2 and G3, as illustrated in Figure 4.1. Cluster G1 is comprised of 9th grade educators, G2 is comprised of 6th grade educators and G3 is a combination of both 6th and 8th grade educators. Figure 4.1 also indicates the presence of bridge nodes in the graph. Bridge nodes represent users who used
resources characteristic of multiple clusters. For example, a 9th grade educator (cluster G1) using resources primarily used by 6th grade educators (Cluster G2). Prior research have found that bridge nodes may serve as important pathways for the transfer of information between disparate groups/clusters in a network [66, 50].

4.4.2.2 Homophily in the deduced network

I now proceed with my second task in understanding the nature of the social network. In this task, I explored the relationship between tie strength and homophily in the deduced social network.

In describing the strength of ties in a social structure, Granovetter posits that the strength of a tie is directly related to the level of homophily between the individuals it connects [66]. In this setting, the strength of a deduced tie is delineated by the edge weight between a pair of nodes (educators). For a pair of nodes $A$ and $B$, an edge weight of $x$ indicates $A$ and $B$ have used $x$ community-contributed resources in common. I used the categorized survey data described earlier to compute the level of homophily between two nodes.

My analysis proceeded by identifying all edges in the deduced social network (Figure 4.1) for which I had survey data for both nodes connected by the edge. I identified 53 out of the 186 unique edges in the network that satisfied this criteria. In this set of 53 edges, edge weights ranged from 1 to 33 with an average and median weight of 4.18 and 2 respectively. Consequently, I split the set of edges into two groups: those with high edge weights and those low edge weights. Edge weights greater than the median weight of 2 were considered as high edge weights and edge weights less than or equal to the median as low edge weights. In this study, there were 31 low edge weights and 22 high edge weights connecting node pairs.

Using the survey item categories: class size, years teaching, class needs, level of comfort and use of technology when teaching, perceived level of isolation and propensity to use and share resource, I investigated the similarity between educators connected by high edge
weights (weights > 2) versus those connected by low edge weights (edges of weight ≤ 2). T-tests were performed to understand if any significant differences existed between both groups: educators that share a high edge weight and educators that share a low edge weight, for the set of categories in Table 4.1. The hypothesis for each category \( x \) is as follows: “Lower edge weights indicate a higher difference in agreement between connected users on \( x \).” The null hypothesis is that there are no significant differences in the level of agreement on \( x \) between the low-edge weight and high-edge weight groups.

Significant results were found for the **propensity to use and share resources** \((t = 3.052, P = 0.00167 \text{ at } p < 0.05)\) and the **level of comfort and use of technology** \((t = 1.871, P = 0.033 \text{ at } p < 0.05)\) categories. The results of the other four categories: class size, years teaching, class needs, and perceived level of isolation were not significant, and thus are not reported here.

These results suggest that the strength of a tie is directly related to the level of homophily between individuals it connects. Higher edge weights between educators indicate similar levels in their propensity to use and share resources, and also in their level of comfort and use of technology. Furthermore, this result provides preliminary support for the claim that the edges in the deduced social network do constitute weak ties, as described by Granovetter’s theory.

### 4.4.3 Analysis 3: Evolution of the deduced social network

Here I address the third research question of this study on whether the deduced social network follows the triadic closure process as predicted by Granovetter’s theory. As stated in section 2.1.2, triadic closures are a basic group phenomena in social networks that provide explanations for the formation of new edges (connections) in a network. Research shows that in a sufficiently large network, edges that create a triadic closure (e.g. the edge between \( B \) and \( C \) in figure 4.2(d)) are also likely to be bridges. For example, the edge \( B - C \) in figure 4.2(d) may act as a bridge between the communities of \( B \) and \( C \) [50].
The analysis performed was carried out in two steps. First, I determined whether the triadic closures observed in the deduced social network were random. Second, if the occurrence of triadic closures are not random, then developing a computational model for predicting the formation of these closures over time is plausible.

In an undirected graph such as the DSN, triads can be in one of four states (see Fig 4.2):

- Empty: No edges exist between any pair of nodes i.e., \(\neg\exists e_{ij} \forall v_i, v_j\) where \(e\) and \(v\) represent edges and vertices (nodes) in the graph respectively.
- Single edge: Exactly one edge exists in the triad.
- Candidate triad: Two edges exist in the triad.
- Closed: A closed triad is one in which all possible edges in the triad exist i.e., \(\forall v_i, v_j \in V\), there exists \(e_{ij} \in E\) where \(E\) and \(V\) are the set of edges and vertices (nodes) in the triad.

![Figure 4.2: Depiction of the four possible states of a triad in our network](image)

I performed a triadic census on a set of 100 random undirected graphs—with the same density and number of nodes as the deduced social network. These graphs were generated using the Erdos-Renyi process \(^1\) to determine if triadic closures observed in our network can

---

\(^1\) Erdos-Renyi is one of two common mathematical models for generating random graphs [52].
be explained by random processes. A triadic census is a count of all possible triad states available in a network. As indicated in figure 4.3, there is a significant disparity between the triadic census of the deduced social network and the average triadic census of all 100 random graphs. This indicates that the triadic closures in the DSN are not the result of a random process.

Given this finding, I developed a computational model to predict the formation of triadic closures as follows. I mined all candidate triads that occur in the network on a monthly basis to understand what features $F$ of a candidate triad $\triad{F}$ at time $t_i$ best predict whether it will close at time $t_{i+1}$ in the future. My model only considers if a candidate triad will close and not when it closes.

The dataset of triads is comprised of two classes: candidate triads that closed and candidate triads that remained open. For each candidate triad that closed, I captured the state of its features (see Table 4.2) just before closing. If the triad remained open, I captured the state of its features at the end of the observation period. The dataset comprised of 640

---

**Figure 4.3**: Comparison between triadic formations in random graphs versus the deduced network

---
candidate triads. 115 of these triads eventually closed and the other 525 remained open.

Consequently, I performed a classification task based on the set of features of each candidate triad and its eventual state (closed or open) to determine what features best predict whether a candidate triad will remain open or closed. The features $F$ of a candidate triad are inspired by prior research which indicate that the structural properties of existing edges and the common node of a candidate triad are good predictors of triadic closure [74].

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge weight 1</td>
<td>The weight of one of two edges that exist between a candidate node and the common neighbor (e.g. the edge $A - B$ in figure 4.2(c)).</td>
</tr>
<tr>
<td>Edge weight 2</td>
<td>The weight of one of two edges that exist between a candidate node and the common neighbor (e.g. the edge between node $A - C$ in figure 4.2(c)).</td>
</tr>
<tr>
<td>Average edge weight</td>
<td>The average of edge weights 1 &amp; 2.</td>
</tr>
<tr>
<td>Common neighbor degree</td>
<td>The degree of the common neighbor $A$ of nodes $B$ &amp; $C$. The degree of a node is the number of edges connected to it.</td>
</tr>
<tr>
<td>Common neighbor between-</td>
<td>Betweenness centrality of the common neighbor $A$ of nodes $B$ &amp; $C$. The betweenness centrality of a node is the number of shortest paths between any two nodes in the network that passes through it.</td>
</tr>
</tbody>
</table>

To determine what features of a candidate triad best predict the likelihood of its closing or remaining open, I performed the following classification tasks:

- Classification with all features listed in Table 4.2
- Classification with the weights of both existing edges (the first two features highlighted in Table 4.2)
- Classification with the most predictive features.
Using Weka’s CfsSubsetEval attribute selection function, I discovered three features to be most predictive: the average edge weight, common neighbor degree, and common neighbor betweenness centrality. CfsSubsetEval evaluates a set of features by considering the predictive power of each feature along with the degree of redundancy between them. It favors features that are highly correlated with the class while having low correlation with each other [71].

I compared a series of classifiers—Naive Bayes, SVM, C4.5 decision trees and REP tree—to the majority class baseline and found the C4.5 decision trees to have the best performance. The C4.5 decision tree classifier had a 15.8% boost in performance over the majority class baseline using the most predictive features (best) described above. The results of our classification experiments are shown in Table 4.3. All classification experiments were evaluated using a standard 10-fold cross validation.

Table 4.3: Triadic closure prediction performance using J48 Decision trees.

<table>
<thead>
<tr>
<th>Features</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority class baseline</td>
<td>0.673</td>
<td>0.82</td>
<td>0.739</td>
<td>0.82</td>
</tr>
<tr>
<td>All features</td>
<td>0.968</td>
<td>0.969</td>
<td>0.968</td>
<td>0.968</td>
</tr>
<tr>
<td>Edge weights</td>
<td>0.825</td>
<td>0.844</td>
<td>0.828</td>
<td>0.844</td>
</tr>
<tr>
<td>Best features</td>
<td>0.978</td>
<td>0.978</td>
<td>0.978</td>
<td>0.978</td>
</tr>
</tbody>
</table>

These experiments show that triadic closures in the network are not random and can be predicted with a high degree of accuracy. These findings further support the hypothesis that the deduced relationship between educators using community-contributed resources constitute weak ties as they exhibit two core properties: homophily and triadic closures [66].

4.5 Discussion

This study questioned whether the edges in a deduced social network depicting the resource usage patterns of educators constituted weak ties. The results of analyses 2 & 3 both provide confirmatory evidence for this hypothesis. In analysis 2, I showed that the
strength of a tie is directly related to the level of homophily between the educators it connects. In the DSN, educators connected stronger ties, i.e., with an edge weight greater than the median, were similar in their propensity to use and share resources and their reported level of comfort and use of technology when teaching.

In the third analysis, I demonstrated that it was possible to predict the formation of triadic closures with a high degree of accuracy using three features of a candidate triad: average edge weight, common neighbor degree and common neighbor betweenness centrality. Within a social network, triadic closures represent new ties between people and thus new paths for the sharing of information. As stated earlier, in large graphs this property leads to bridge nodes i.e., people that connect sub-communities. Even though the size of the DSN is small, the second study indicated that several bridges nodes have emerged that link communities of educators across different grade levels (see figure 4.1).

4.5.1 Implications for school districts and educators

My first analysis found that educators with higher perceptions of isolation sought out resources contributed by peers in their school district. Other school districts may use this finding as motivation for providing educators with greater access to, and awareness of, resources available through educational digital libraries and online professional learning communities. This research provides preliminary evidence that such repositories may help educators to overcome isolation. Several studies have shown that teacher isolation can have a detrimental effect on the quality of teaching instruction and student learning [34, 53, 130]. In the school district I studied, many of these educators are the only Earth science teacher in their school. The CCS provided these teachers with one of the few means by which they could view and share Earth science materials created by their peers. While I only examined educators’ perceived level of isolation prior to system deployment, future research should investigate whether prolonged system usage helps to alleviate perceptions of isolation.
4.5.2 Benefits of mixed methods research

In these studies, I combined network analytic techniques with social science research methods such as surveys. This mixed methods approach was essential: the network analytic techniques uncovered interesting phenomena, while the survey data provided insight into the meaning of these phenomena. For instance, the survey data was essential to identifying the similarities between educators in the DSN (homophily). Additionally, the survey data enabled us to understand the social-psychological constructs influencing educator use of community-contributed resources, which in this case is educator isolation. In bounded institutional settings, mixed methods approach can help to identify the specific social and psychological concerns that shape system use and the evolution of the institution’s social network.

4.6 Limitations

In comparison to contemporary social network research, the datasets analyzed in this study are small. Consequently, I have taken steps to verify that the results being reported are statistically significant. The size of the datasets are a direct function of the research context: patterns of sharing among educators in a single school district. School districts, like many companies, are very interested in learning how the tools and processes they have put into place support professional learning and knowledge sharing amongst their employees. These types of network analysis focusing on activities within a bounded institutional setting, will often have to face the challenges associated with smaller datasets. Nevertheless, I believe that it is imperative to develop data analytic techniques that are appropriate to studying small communities within institutional settings.
4.7 Conclusion

In this study, I showed that Granovetter’s theory on weak ties can be used to understand the phenomena of homophily and triadic closures in the deduced social network. Furthermore, I discovered that an educator’s perceived level of isolation in the classroom may play an influential role in their decision to use community-contributed resources.

To attain these insights, I combined network analytic techniques with social science research methods such as surveys. This mixed methods approach was essential: the network analytic techniques uncovered interesting phenomena, while the survey data provided insight into the meaning of these phenomena. For instance, the survey data was essential to identifying the similarities between educators in the network (homophily). Additionally, the survey data illuminated the social-psychological constructs which may influence the formation and evolution of the deduced social network. I believe that the methods described in this study can be generalized to other bounded institutional settings to identify the specific social and psychological concerns that shape system use and the institution’s social network.
Chapter 5

Using network analysis of clickstream data to improve resource recommendation systems

5.1 Purpose

The purpose of the study presented in this chapter is to determine if the triadic closure prediction model developed in the first study (see chapter 4) can be used to improve traditional resource recommendations systems. In particular, this study investigates the extent to which the triadic closure prediction model developed in chapter 4 can be used to improve traditional collaborative filtering recommendation systems.

5.2 Data

This study is based on the recorded clickstream of Earth Science educators who used the CCS during the 2011-2012 and 2012-2013 academic years. It uses the triadic closure prediction model developed in study 1 (see chapter 4) to augment recommendation systems that were built and evaluated on clickstream data from the 2012-2013 school year. As a reminder, the triadic closure prediction model was developed using clickstream data from the 2011-2012 school year. The 2011-2012 and 2012-2013 datasets are described in Table 5.1.
Table 5.1: Description of datasets under analysis. The 2011-2012 data is used in building the triadic closure prediction model (see chapter 4), and the recommendation systems are built and evaluated on the 2012-2013 data.

<table>
<thead>
<tr>
<th>Stat</th>
<th>School year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2011-2012</td>
</tr>
<tr>
<td>Number of users</td>
<td>40</td>
</tr>
<tr>
<td>Number of schools</td>
<td>31</td>
</tr>
<tr>
<td>Grades</td>
<td>6th-9th</td>
</tr>
<tr>
<td>Number of community resources used</td>
<td>372</td>
</tr>
</tbody>
</table>

5.3 Methodology

The recommendation engine is developed in two stages. In the first stage, collaborative filtering and content-based recommendation systems were implemented separately. They were then combined in a hybrid recommendation system.

In the second stage, the triadic closure prediction model was added to the collaborative filtering system and then to the hybrid system to gauge whether the addition of the triadic closure prediction model could improve the accuracy of either recommendation system. The triadic closure prediction model was not applied to the content-based system because content-based recommendations are by nature, independent of user similarity.

I now describe how each of the aforementioned recommendation systems were implemented.

5.3.1 Collaborative filtering resource recommendation system

The collaborative filtering recommendation system recommends to a user $u$ resources that have been utilized by other users similar to $u$. Two users $u, v$ are judged to be similar based on their ratings for a set of resources used between them. The CCS click stream dataset consists of only unary ratings on resources. Specifically, we only know which resources an educator used (clicked on). For the set of resources used by users $u$ and $v$, user similarity is determined by computing the Jaccard similarity between $u_r$ and $v_r$ where $u_r$ is the set of
resources used by user \( u \) and \( v_r \) is the set of resources used by user \( v \):

\[
sim(u, v) = \frac{|u_r \cap v_r|}{|u_r \cup v_r|}
\]  

(5.1)

The Jaccard similarity of two sets \( u_r \) and \( v_r \) is the ratio of the size of the intersection and union between \( u_r \) and \( v_r \) \cite{107} as shown in equation 5.1. The Jaccard similarity between two users is a real value ranging from 0-1. Two users are similar if the set of resources they have used has a high Jaccard similarity. In this study, I deem users with a Jaccard similarity score of \( \geq 0.1 \) to a user \( u \) as similar enough to be considered for a collaborative-filtering recommendation of resources to \( u \).

5.3.2 Content-based resource recommendation system

The content-based recommendation system recommends to a given user \( u \), resources that are similar to the set of resources that have already been used by \( u \). Resource similarity is computed based on a set of seven features. These features and their weights are highlighted in Table 5.2. The first six of these features (resource title, description, context, author, grade tags, and other tags) are surface features i.e. they are visible to all users of the CCS platform (see Figure 1). The 7th feature (the resource type) was created by grouping resources according to their intended purpose. For example, resources contributed by users for use in testing and assessment purposes were grouped as assessments, and those created as reading aids for students were grouped as reading guides. The similarity between two resources \( b \) and \( c \) is given by the following equation:

\[
sim(b, c) = \sum_{f=1}^{n} Jaccard(b_f, c_f)w_f
\]  

(5.2)

where \( f \) indicates a feature of the resource, and \( w \) indicates the weight of that feature.

As with the collaborative filtering recommendation system, only resources with a similarity score of \( \geq 0.1 \) to a resource \( b \) are considered by the content-based recommendation system in making resource recommendations that are similar to \( b \).
Table 5.2: Description of features of a community resource

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>This refers to the contributor of a resource. Each resource contributor is identified by a unique user id.</td>
<td>20%</td>
</tr>
<tr>
<td>Context</td>
<td>This refers to the specific curricular context within the CCS to which a resource was contributed. This curricular context could be a specific unit or activity. Resources can also be contributed to a global repository called “All shared stuff.” Resources in “All shared stuff” are accessible from all contexts within the CCS.</td>
<td>10%</td>
</tr>
<tr>
<td>Title</td>
<td>This refers to the title of a resource. The content-based recommendation system is operationalized on keywords that are extracted from a resource’s title. I use a rapid keyword extractor (RAKE)(^1) to extract keywords from a resource’s title.</td>
<td>10%</td>
</tr>
<tr>
<td>Description</td>
<td>This is a short blurb describing the resource. Keywords are extracted from the resource’s description and are used in comparing it to other resources. I use a rapid keyword extractor (RAKE)(^2) to extract keywords from a resource’s description.</td>
<td>10%</td>
</tr>
<tr>
<td>Grade-level tags</td>
<td>These are a set of user assigned tags that give an idea of the scope of a resource. For example, a resource with the tags ‘GT’ and ‘ELA’ indicates that this resource is appropriate for both Gifted &amp; Talented students and is appropriate for improving the English literacy of students.</td>
<td>20%</td>
</tr>
<tr>
<td>Other tags</td>
<td>These are a set of user assigned tags that indicate the scope of a resource i.e., the range of instructional contexts that a resource is suitable for. These tags include the appropriateness of a resource for instructing student of special abilities such as gifted and talented students or English language learners.</td>
<td>10%</td>
</tr>
<tr>
<td>Resource type</td>
<td>This refers to the intended purpose of a resource. Community-contributed resources were categorized into one of eight high level categories depending on their purpose. They are: animations, assessments, reading guides, worksheets, flipcharts, lab exercises, web links and power point presentations.</td>
<td>20%</td>
</tr>
</tbody>
</table>

5.3.3 **Hybrid resource recommendation system**

The hybrid recommendation system is a combination of both the content-based and collaborative filtering recommendation systems. This system returns the set of resources produced by the intersection of the results of both the content-based and collaborative filtering
recommendation systems.

5.3.4 Triadic closure augmentations of both the collaborative filtering and hybrid recommendation systems

As discussed earlier, the collaborative filtering recommendation system recommends resources to a user $u_i$ based on resources utilized by other users who are similar to $u_i$. Similarity is computed using the Jaccard similarity metric. Thus if $sim(u_i, u_j) = 0$, resources of $u_i$ would not be recommended to $u_j$ and vice versa. In this domain, $sim(u_i, u_j) = 0$ would occur when there is no edge $e_{ij}$, e.g. the missing edge $B - C$ in Figure 4.2(c). Thus, if the recommendation system predicts that a candidate triad with no edge $e_{ij}$ is likely to close in the future, it can use this knowledge to recommend resources of $u_i$ to $u_j$ even though $u_i$ and $u_j$ are currently not similar.

Once a candidate triad is predicted to close, the similarity of the user (node) in the closure is computed by averaging the similarity between the common neighbor and the two other nodes in the triad. For example, consider the case where a recommendation of resources is being made to the node B in the candidate triad in Figure 4.2(c). Using a traditional collaborative filtering system, only resources of A would be recommended to B since $Jaccard(A_r, B_r) > 0$, where $A_r$ and $B_r$ represent the set of resources used by nodes A and B respectively. When the traditional collaborative filtering system is augmented with the triadic closure prediction model, it makes a prediction as to whether the edge $B - C$ is likely to form. If this prediction is positive, then the similarity of node B to C is computed by taking the average of the similarities of A to B and A to C i.e.

$$sim(B, C) = \frac{Jaccard(A_r + B_r) + Jaccard(A_r + C_r)}{2}$$

where $A_r$, $B_r$ and $C_r$ represent the set of resources used by nodes A, B and C respectively. A similarity score of $\geq 0.1$ is required to recommend resources from C to B as with the collaborative filtering and content-based recommendation systems.
5.3.5 Evaluating recommendation systems

Prior studies have proposed three main ways for evaluating recommendation systems. They are: offline experiments, user studies and online experiments [118]. I now describe each of these evaluation approaches, after which, I investigate which approach is the most suitable to evaluating the recommendation system being developed in this study.

5.3.5.1 Offline experiments

Offline experiments involve testing the output of the recommender system on existing data. The typical measure for offline experiments is prediction accuracy. For a test set $T$ of actual usage, a recommender system predicts a set of ratings for $T$. Offline experiments are widely used in evaluating recommender systems due to the low set up cost and the ability to compare a wide range of different candidate algorithms easily [118]. However, by reducing the recommendation problem to a prediction task, offline tests have been discredited for their inability to measure the novelty or serendipity of recommendations [72]. Novel recommendations are successful recommendations of items the user did not know about and serendipity is a measure of how surprising a successful recommendation is [118]. For example, a successful movie recommendation that introduces a user to a new genre or actor is more serendipitous in comparison to one where the user is already familiar with the acting cast and genre.

5.3.5.2 User studies

User studies involve having a set of users perform pre-determined tasks using the recommendation system. Quantitative and qualitative data such as the amount of time spent on tasks and user perception of the system interface is taken as part of the study [118]. As an example, consider a user study designed to test the impact of a recommender system on the browsing behavior of users on a news site. Users are asked to read a set of
stories on the site that are interesting to them; some, but not all, of these stories are system recommendations. Afterwards, quantitative measurements on the performance of users on the task are taken. These include the number of times the recommendations were clicked on, and tracked gestures such as eye movements that show whether subjects looked at the recommendations. Qualitative data include post-study interviews that examine the relevance of recommendations to users [118]. Qualitative studies investigate the reason behind a user’s decision to select a news story. Unlike offline experiments, user studies allow system designers to explicitly study the behavior of users as they are interacting with the recommender system. However, user studies are expensive to conduct and the outcome of the experiments maybe skewed, depending on the types of participants tested [118]. For example, if there is a significant disparity in the interests and/or browsing behavior of participants in the user study as compared to real users, then insights from the user study may not provide an accurate estimate of general user behaviour.

5.3.6 Online experiments

Online experiments are widely regarded as the most accurate approach for evaluating recommender systems [72]; they evaluate the performance of recommendation systems on real users oblivious to the experiment. Most times, online experiments are a penultimate step in the evaluation of a new recommendation system. They give system designers the best insights into the impact of a recommender system on user behavior [118]. A simple example of an online experiment is an A/B test where a subset of users to a website are shown recommendations. Quantitative data is collected to understand whether there are any substantive differences in the usage behavior and interaction patterns of the subset of users that viewed the recommendations and those that did not [72, 118].
5.3.7 Using offline experiments to evaluate the recommendation systems developed

Given that I only had historic usage data (i.e. the recorded user click stream from the 2012-2013 school year) to build and evaluate the recommendation systems on, I used an offline experiment approach to evaluate the recommendation systems developed. Specifically, the offline experiment was modeled as a prediction task.

The offline experiment was carried out as follows: for each day in which at least one community-contributed resource was used, the recommender system made predictions for the set of resources that were used on that day based on the click stream data up until the day prior. For example, recommendations for August 5th will be generated using clickstream data prior to August 5th. Recommendations are limited to users who have a history of using community-contributed resources i.e. the system does not attempt to predict a set of recommendations for a user’s initial use of community-contributed resources. In the recommender system literature, this is commonly known as the cold-start problem and has garnered a lot of research interest in recent times [117, 79, 113].

The precision at \( N \) [72, 118] was computed for each set of predictions made by the recommendation system. Precision at \( N \) (\( P@n \)) measures the ratio of the number of accurate predictions to the total number of recommendations made where \( N \) represents the total number of recommendations made. It is given by the formula:

\[
Precision = \frac{true - positive}{true - positive + false - positive}
\]

For a user \( u \), true-positives are the recommended resources that were used by \( u \) and false-positives are the resources that were recommended but not used by \( u \).

\( N \) is set as the number of resources used by a user \( u \). Thus, if the system makes predictions for a \( u \)’s usage of 5 resources on August 5th, then \( N \) will be the recommendation system’s top 5 recommendations. Other studies have also used recall in measuring the performance of recommender systems. However, recall is generally not a practical evaluation
measure for recommender systems since relevance is determined at the user level [72]. In traditional Information-Retrival (IR) experiments, relevance is standardized at the query level, i.e. for a particular query, we can objectively determine the relevance of the returned documents to that query. However, in recommendation systems, relevance is an individual metric. A resource/item that is relevant to one person may be irrelevant to another person. Furthermore, we have no way of knowing whether a resource was not used because it is not irrelevant to the user. Even if we say that the usage of a resource equates to relevance, then for a fixed value of $N$, Recall at $N$ is effectively the same as Precision at $N$. Thus calculating Recall at $N$ will be effectively be the same as calculating Precision at $N$.

Consequently, I calculated the average precision at $N$ for the five recommendation systems developed:

(1) Collaborative filtering

(2) Content-based

(3) Collaborative-filtering + triadic closure prediction model

(4) Hybrid system (collaborative-filtering + content-based)

(5) Hybrid system + triadic closure prediction model.

The results of the experiments carried out to evaluate the recommendation systems listed above are highlighted in Table 5.3

Table 5.3: Average Precision at $N$ for all five recommendation systems evaluated.

<table>
<thead>
<tr>
<th>Recommendation system</th>
<th>Average precision at $N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborative filtering</td>
<td>17.26%</td>
</tr>
<tr>
<td>Collaborative filtering + Triadic closure prediction model</td>
<td>29.12%</td>
</tr>
<tr>
<td>Content-based</td>
<td>68.57%</td>
</tr>
<tr>
<td>Hybrid system (Collaborative filtering + Content-based)</td>
<td>77.15%</td>
</tr>
<tr>
<td>Hybrid system + Triadic closure prediction model</td>
<td>83.25%</td>
</tr>
</tbody>
</table>
The results show that the triadic closure prediction model improves by 68.66% (17.25% - 29.12%) the average precision at $N$ of a traditional collaborative filtering recommendation system. The triadic closure prediction model also improves by 7.9% (77.15% - 83.25%) a traditional hybrid recommendation system. Paired t-tests indicate that the improvements of the triadic closure prediction model over a standard collaborative filtering $(t = 3.368, P = 0.000417)$ and hybrid recommendation system $(t = 1.744, P = 0.041)$ are statistically significant at $p < 0.05$.

Also, from Table 5.3, we can see that the content-based recommendation system had the greatest impact on overall recommendation accuracy. There are two main explanations for this.

First, the deduced social network of educators (see Figure 4.1) is primarily characterized by the grade levels they teach. Thus, most of the 6th grade educators primarily used 6th grade resources, and the 9th grade educators primarily used 9th grade resources. Also, most of the resources used by educators during the 2012-2013 school year had been contributed to the system in previous years; thus, the content-based system had a large repository of possible resources to recommend to users. This improved the probability of the top $N$ recommendations matching the set of resources that were used by users.

Second, the collaborative filtering model (and triadic closure model) recommendations suffered early on in the school year due to the paucity of usage click stream data with which to make recommendations. Since the academic school year did not fully begin until late August, very few users utilized the system in the month of August, thus making it difficult for the collaborative filtering recommendation system to find similar users from which to make recommendations for the current user. This follows findings in the literature that show collaborative-filtering systems suffer when the user-item matrix is very sparse [3, 51, 107].
5.4 Discussion

In this study, I augmented a traditional collaborative filtering recommendation system with the triadic closure prediction model from the first study. Preliminary evaluations of this approach were promising as it outperformed both a traditional collaborative filtering and a hybrid recommender system. I now discuss potential implications of my findings with regards to trust-based recommendations and I also discuss the limitations of this approach in its current state.

5.4.1 Triadic closures and trustworthy recommendations

The triadic closure prediction model may represent a novel way of incorporating elements of trust-based recommendations within a deduced social network. In the first study, I discovered that the edge weight between two educators is directly proportional to their level of homophily (similarity). In particular, I discovered that higher edge weights between educators indicate greater similarity in their propensity to use and to share resources, and greater similarity in their level of comfort and use of technology.

Consequently, as discussed in Section 4.4.3, the average edge weight between two nodes $B, C$ and their common neighbor $A$ (see Figure 4.2(c)) is indicative of a triadic closure. Prior studies has illustrated the direct relationship between homophily and trust [94], and more recent work have linked trust to the efficacy of a recommendation [85]. Individuals are more likely to accept recommendations from people they trust [12]. However in networks with sparse connections between users, pure trust-based recommendation systems suffer [90]. To alleviate this drawback, trust propagation mechanisms have been proposed [12] to improve recommendation quality. These approaches focus on propagating trust across relationships and are similar to our triadic closure prediction model [90]. Thus if user $A$ trusts $B$ and $C$ trusts $B$, trust propagation mechanisms assume that to some extent, $B$ is also likely to trust $C$. 
5.4.2 Limitations

There are some drawbacks to the triadic closure recommendation system as outlined in this study. The biggest disadvantage of the triadic closure prediction model is its computational complexity. For each point in time when a recommendation is to be made, two computationally expensive operations are performed. First, a deduced social network of the click stream prior to that point has to be generated. Second, the deduced network has to be scanned to detect all candidate triads (see Figure 4.2(c)), each of which is evaluated for the possibility of closing at time $t_{i+1}$. Generating the deduced network scales linearly as the size of the network and detecting all candidate triads is an $O(n^2)$ operation. Therefore, in a very large network it may not be computationally feasible to perform both of these operations in real time. Similarly, the naive collaborative-filtering implementation illustrated in this paper does not scale in applications with a very large user-item matrix, such as applications with millions of users and items/resources [86]. A quick optimization to the triadic closure prediction model will be to implement the algorithm using a dynamic programming-like approach. Dynamic programming works by caching intermediate results that tend to be performed multiple times by related sub problems. In this setting, this could be caching the results of both operations listed above. Thus, in generating the deduced network, the recommendation system would only generate new parts of the network that change from a previous run of the algorithm, and in predicting the formation of triadic closures would only scan parts of the network that have changed over time.

Furthermore, techniques for improving collaborative-filtering and content-based recommendation approaches were not explored in this research. These include dimensionality reduction—i.e. reducing the user-item matrix—and grouping related items in the user-item matrix have not be explored. Although these techniques would have improved the computational complexity of the collaborative-filtering and content-based approaches, they may not have accounted for the edge cases discovered by the triadic closure prediction model.
Future work incorporating and measuring the impact of these optimizations in improving the recommender system will make for an excellent line of inquiry.

5.5 Conclusion

Recommendations are a staple of modern e-commerce, content and information delivery platforms. Within learning environments, appropriate recommendations are necessary to keep learners informed and engaged; and educators abreast of appropriate pedagogical resources for classroom instruction.

In this study, I introduced an approach that improves on a standard hybrid recommendation system approach by performing network analysis on a deduced social network generated from user clickstream data. In particular, this approach predicts the formation of triadic closures in a deduced social network and uses this prediction to augment a traditional collaborative filtering approach. The triadic closure prediction model is able to predict the formation of edges between users in a deduced social network—thus future interest in the same resources—where a traditional collaborative filtering recommendation approach cannot. My model shows a 68.66% and 7.9% boost to the average precision at $N$ of a traditional collaborative filtering and hybrid recommender system respectively. These improvements in recommender performance were found to be statistically significant.

I profer the approach outlined in this study as only a first step towards improving the quality of a recommender system. Before deploying a recommendation system, a more thorough evaluation of its quality by experts in the learning domain is suggested. These could include controlled user studies or online experiments. Current literature indicates that these evaluation approaches are the most indicative of a recommender system’s impact on user behavior [72, 118].
Chapter 6

An information cascade model for understanding the usage distribution of resources

6.1 Purpose

This study investigates the usage distribution of community-contributed resources in the CCS. The usage of a community-contributed resource is defined by the number of unique educators who use (click on) it. These resources include lesson plans, videos, reading guides and other types of instructional materials that educators have found useful in their instruction, and subsequently shared with their peers. I explored what the usage distribution of these resources looked like and investigated what underlying mechanisms may have generated the observed distribution.

6.2 Data

This study is based on the click stream of 6th and 9th grade Earth Science educators who shared and used community-contributed resources in the CCS over a period of four academic years: 2009-2010, 2010-2011, 2011-2012, and 2012-2013. A core design goal of the CCS is to enable educators to share their instructional practices with others in their district through these community contributions. The datasets I analyzed are summarized in Table 6.1.
Table 6.1: Description of clickstream datasets under analysis. NA indicates that this statistic does not apply to the year under observation

<table>
<thead>
<tr>
<th>Stat</th>
<th>09-10</th>
<th>10-11</th>
<th>11-12</th>
<th>12-13</th>
<th>All years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
<td>89</td>
<td>50</td>
<td>40</td>
<td>55</td>
<td>154</td>
</tr>
<tr>
<td>Number of new users in this school year</td>
<td>89</td>
<td>16</td>
<td>23</td>
<td>26</td>
<td>NA*</td>
</tr>
<tr>
<td>Number of resources used</td>
<td>386</td>
<td>249</td>
<td>372</td>
<td>325</td>
<td>532</td>
</tr>
<tr>
<td>Max number of users per resource</td>
<td>24</td>
<td>20</td>
<td>14</td>
<td>9</td>
<td>44</td>
</tr>
<tr>
<td>Number of resources with only 1 user</td>
<td>25</td>
<td>89</td>
<td>209</td>
<td>195</td>
<td>77</td>
</tr>
<tr>
<td>Mean number of users per resource</td>
<td>5.66</td>
<td>2.52</td>
<td>1.79</td>
<td>1.68</td>
<td>7.297</td>
</tr>
<tr>
<td>Median number of users per resource</td>
<td>5.00</td>
<td>2.00</td>
<td>1.0</td>
<td>1.0</td>
<td>6.00</td>
</tr>
<tr>
<td>Number of new resources used over prior year(s)</td>
<td>386</td>
<td>37</td>
<td>50</td>
<td>59</td>
<td>NA*</td>
</tr>
</tbody>
</table>

6.3 Methodology

This study began with an exploration of the usage distribution of community-contributed resources. Specifically, I was interested in finding out whether the observed usage distribution followed any known probability distribution. Consequently, I explored the underlying mechanism(s) that may have given rise to the observed distribution. Specifically, I considered whether resource position or perceived quality could predict the usage distribution. Through the lens of information cascades, I also questioned the extent to which social influence impacted the observed usage distribution. I now describe each of the analyses performed as part of this study.

6.4 The usage distribution of community-contributed resources

Visually, the usage distribution of community-contributed resources during each academic year and across all academic years appears right-skewed and heavy-tailed, as shown in Figures 6.1 and 6.2. However, the distribution of the 2009-2010 academic year appears to be
more uniform in comparison to that of subsequent years. To gain a more quantitative and concrete understanding of the differences between the usage distribution of each academic year, I looked into the skewness and kurtosis of the usage distribution of each academic year.

Skewness is a measure of symmetry in a distribution. A distribution is symmetric if it looks the same to the left and right of the center point. The skewness of a distribution is defined by the following formula $\gamma = \mu_3/\mu_2^{3/2}$ where $\mu_2$ and $\mu_3$ are the second and third central moments. My analyses indicate that the usage distribution of each academic year is positively skewed (see Table is 6.2). However, the distribution of the 2010-2011, 2011-2012 and 2012-2013 academic years are considerably more positively skewed than the usage distribution of 2009-2010.

While skewness measures the symmetry of a distribution, kurtosis measures a distribution’s peakedness. Kurtosis gives an idea of how heavy-tailed a distribution is relative to

---

1 Moments are a set of statistical measures used to describe a distribution. For example, the first and second central moments of a distribution are its mean and variance.
the normal distribution. It is given by the following formula: \( \gamma_2 = \mu^4 / \mu_2^2 - 3 \) where \( \mu_2 \) and \( \mu_4 \) are the second and fourth central moments. Distributions with high kurtosis tend to have a distinctive high peak near the mean, followed by a sharp decline and long tail. Distributions with low kurtosis, such as the uniform distribution, tend to be flat. Table 6.2 shows that the usage distributions of every year—apart from the 2009-2010 academic year—have a high kurtosis. In fact, the distributions of the 2010-2011 and 2011-2012 academic years have about 10 times more kurtosis than the 2009-2010 academic year. Also, the 2012-2013 academic year has about seven times more kurtosis than that of the 2009-2010 academic year.

Given the skewness and kurtosis of the usage distribution of community-contributed resources in the CCS, it is unlikely for the usage of these resources to be normally distributed. However, to conclusively rule out the normal distribution, I performed a Shapiro-Wilks test for normality on the distribution. The null hypothesis \( H_0 \) of the Shapiro-Wilks test is that the data is normally distributed and the alternate hypothesis \( H_a \) is that the data is not normally distributed. The outcome of the Shapiro-Wilks test showed that the null hypothesis \( H_0 \) can be rejected with a p-value of \( 2.2 \times 10^{-16} \). This provides conclusive evidence that the usage of community-contributed resources is not normally distributed.

Table 6.2: Comparison of the skewness and kurtosis of the usage distribution for each academic year

<table>
<thead>
<tr>
<th>Measure</th>
<th>09-10</th>
<th>10-11</th>
<th>11-12</th>
<th>12-13</th>
<th>All years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>1.1675</td>
<td>3.776</td>
<td>3.528</td>
<td>3.156</td>
<td>1.704</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>2.269</td>
<td>23.485</td>
<td>22.491</td>
<td>14.347</td>
<td>5.709194</td>
</tr>
</tbody>
</table>

6.4.1 User behavior and the observed heavy-tailed distributions

Furthermore, I investigated whether the usage behaviors of educators gave any insights into the skewed distribution in the usage of resources. In particular, I investigated whether

---

2 A standard statistical test for exploring the normality of data
there was a contrast between the number of resources used (clicked on) before they had been saved (i.e. had a public social signal) versus those that were used only after they had already been saved. The outcome of this analysis was quite startling, as shown in Figure 6.3. During the 2009-2010 academic year, only 14.5% of all resources used were clicked on before they had been saved. That is, only 14.5% of all resources used had no public social signal prior to usage. This number drops to zero during the 2010-2011 and 2011-2012 academic years. Only 2 out of the 325 resources that were used during the 2012-2013 academic year had not been saved by a user prior to usage.

This preliminary analysis suggests that social influence may have played a role in the observed skewed distribution. It gives a precursory indication that educators were more likely to use resources that had already been saved by others.

### 6.4.2 Usage distribution of community-contributed resources as a power law

My next analysis investigated whether the observed heavy-tailed shape of the usage distribution of community-contributed resources followed any known probability distribution.
Figure 6.3: A comparison between the total number of resources used and the number of resources used before they had been saved per academic year.

Specifically, I investigated whether it followed a power law. Since power laws are known to experience the rich-get-richer phenomenon—i.e. entities that are popular tend to get even more popular over time—power laws provide a logical hypothesis for the observed distribution. One can imagine that community-contributed resources with more saves, hence visibly more popular, will attract even more usage over time.

Determining whether a distribution exhibits a power law is a complex task [100]. Typically, investigations begin with a log-log plot of a quantity’s complimentary cumulative distribution function (Pr[X ≥ x]). A resultant straight-line plot provides a preliminary indication of a power law. I generated a log-log plot of the complimentary cumulative distribution function (CCDF) of the usage of each resource, i.e. the number of users who used each resource. A straight line—the slope of which represents the scaling parameter α—was observed, as shown in Figure 6.4. This provided preliminary evidence for the power law
hypothesis.

While historically, power laws have determined from a visual appraisal of a quantity’s CCDF, visual evidence is not sufficient enough to ascertain a power law. Recent research indicates that qualitative determinations of power laws from plots of logarithmic transformations of data are fraught with errors [38, 100]. Data from other probability distributions such as the lognormal or exponential distribution can also show up as a straight line on a log-log plot [100, 38]. Thus, to garner plausible evidence for the power law hypothesis, I follow a principled statistical approach introduced by Clauset et al. [38] for identifying power law distributions in empirical data. Clauset et al. [38] prescribe a three step approach for determining power laws in empirical data. They are as follows:

1. Determine the values of $x_{min}$ and the scaling parameter $\alpha$ of the data that makes the probability distribution of the measured data and best-fit power law model as
close as possible for values of $x$ where $x \geq x_{min}$. This is done using a Maximum Likelihood Estimator (MLE) approach that determines the values of $x_{min}$ and $\alpha$ for which the Kolmogorov-Smirnov (KS) statistic (distance) between the empirical data and power law model is smallest. The KS test is a non-parametric test for determining if two data distributions $d_1$ and $d_2$ differ significantly [91]. It returns a value $D$ which specifies the distance between $d_1$ and $d_2$.

(2) Perform a goodness of fit test between the data and synthetic datasets drawn from power law distributions parameterized by the values of $x_{min}$ and $\alpha$ from step 1. Typically, 2500 synthetic datasets are generated from a power law distribution parameterized by $x_{min}$ and $\alpha$. The premise for generating 2500 synthetic datasets is simple. If we wish to obtain statistically significant results within about $\epsilon$ of the true value, then we should generate at least $\frac{1}{4}\epsilon^{-2}$ synthetic datasets [38]. If we wish our p-value to be accurate to within 2 decimal digits, then $\epsilon = 0.01$, and thus therefore, 2500 synthetic datasets will be needed since $\frac{1}{4}(0.01^{-2}) = 2500$. A p-value is generated by calculating the fraction of the synthetic datasets with a higher KS distance to the power law fit in comparison to the empirical data. When this value is closer to 1, then the differences between the empirical data and the model can be attributed to statistical fluctuations alone and we can conclude that the model is a plausible fit to the data [38]. Clauset et al. [38] suggest a value of $p > 0.1$ to reasonably accept the power law hypothesis as a plausible fit to the data.

(3) Perform likelihood ratio tests to determine if alternative distributions provide a better fit to the data. Clauset et al. [38] encourage investigators to at least compare the fit of their data to exponential, log normal and stretched exponential (Weibull) distributions. If any other probability distribution has a statistically significant better fit compared to power law distribution, the power law hypothesis is rejected in favor of the alternate distribution.
I used software packages [58, 8] implementing the approaches of Clauset et al. [38] to determine if the observed usage data follows a power law.

Consequently, the usage distribution of community-contributed resources across all academic years was found to follow a power law with an $\alpha$ of 4.44 and an $x_{min}$ value of 15. Goodness of fit tests (step 2) determined that this fit had a p-value of 0.86. Furthermore, likelihood ratio tests (step 3) comparing this fit to a log normal, exponential and Weibull distribution indicated that none of the alternatives provided a significantly better fit to the empirical data. Figure 6.5 illustrates that a power law provides a closer fit to the complimentary cumulative distribution function (CCDF) of the empirical data in comparison to the lognormal and exponential distribution.

Figure 6.5: Comparisons of the complimentary cumulative distribution function (CCDF) of the empirical data, the power law, lognormal and exponential distribution fits to the data.
6.5 Mechanisms behind the power law distribution of resources

My next order of investigations looked into generative mechanisms that might have been responsible for the observed power law distribution in the usage of community-contributed resources. I considered if resource position, quality or social influence—investigated through the lens of information cascades—can provide some explanation as to how the observed power law distribution arose.

6.5.1 Resource position

My first line of inquiry looked into whether the position of a resource played any role in its usage. I began by performing simple correlation tests between the position of resources and their usage. If resource usage is influenced by position, then we would expect a negative correlation. That is, resources at lower positions (top of the list) will have greater usage (number of users). In the CCS, resources are organized in lists ordered from positions 0-9. Position 0 represents the top of the list and 9 is the bottom of the list. Most resources are accessible via multiple lists in the CCS. Thus, the recorded click position of a resource is dependent on which list it is located in. In light of this, correlation tests were performed against multiple modalities of a resource’s position namely: the median click position, the last click position and the mode click position. The median click position refers to the middle position of all recorded clicks on the resource. The mode click position refers to the most frequent position of all clicks on the resource. The last click position refers to the position of the last recorded click on the resource. The null hypothesis $H_0$ of the correlation tests is that there is no correlation between the click position of a resource and the number of users who clicked on it. The alternate hypothesis, $H_a$ is that the true correlation is less than 0 i.e. the lower the position of a resource (higher rank), the greater the number of users.

As shown in Table 6.3, results indicate only a weak correlation between the mode click position and resource usage during the 2012-2013 academic year. This strongly suggests that
the position of a resource is not responsible for the observed power law distribution.

Table 6.3: Correlation tests between the last, mode and median click position of resources and their eventual usage. Statistically significant correlations are highlighted in bold text.

<table>
<thead>
<tr>
<th>Stat</th>
<th>School year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>09-10</td>
</tr>
<tr>
<td>Last click</td>
<td>0.0216</td>
</tr>
<tr>
<td>Mode</td>
<td>-0.0247</td>
</tr>
<tr>
<td>Median</td>
<td>-0.0345</td>
</tr>
</tbody>
</table>

6.5.2 Resource quality

My next analysis investigated if any relationship existed between the quality of a resource that can be inferred before a user clicks on it and its usage. This analysis is based on all resources used between 2009 and 2013 (see Table 6.1). I was able to extract the records of 523 out of the 532 resources used within this period. The xml annotation record of the remaining 9 of these resources were unavailable at the time of this analysis. My analyses were carried out in two steps:

- First, I used the presence of a description in the listing of a resource as a marker of its quality. Thus, resources with a description were deemed as having high quality and those without a description were of low quality. I then investigated if there was a statistically significant difference in usage between resources of high quality and those of low quality. Since user contributed, educators have the option of writing a description along with the resource they contribute. However, not all educators do so.

- Second, I mapped quality indicators of Bethard et al. [17] to signals that can be inferred by users before using a resource.
6.5.2.1 Differences between resources with a description and resources with no description

To determine if any significant differences in usage existed between resources with a description ($n = 276$) and those without one ($n = 247$), I performed a Mann-Whitney-Wilcoxon test between both groups. The null hypothesis $H_0$ is that the usage distribution of both groups is the same and the alternative hypothesis $H_a$ is that resources with a description have a greater market share than resources without a description. I was unable to reject the null hypothesis (p-value: 0.1665). This indicates that there was not a statistically significant difference in the usage of resources with a description and those without a description.

6.5.2.2 Resource usage and quality indicators

My next analysis investigated whether there was any correlation between the number of inferable quality signals of a resource and its usage. To do this, I developed a composite resource quality score that incorporated all signals of a resource’s quality that can be inferred by a user before using it. These signals were mapped to the resource quality indicators developed by Bethard et al. [17] as described in Table 6.4. The resource quality score is a real number between $0−1$. The weights of each of these signals contributing to a resource’s quality score is described in Table 6.5.2.2. They were determined from the number of keywords in a resource’s description and the presence of time and grade tags. If a description had $\geq 8$ keywords, it received the full weight for the description signal. Else, if $4 > $ keywords $< 8$, $1/2$ the weight for the description signal was given. If the resource had $< 4$ keywords then no weight was assigned to the description signal. The full weight of time tag and grade tags required the presence of at least 1 time tag and 1 grade tag respectively.

A correlation test was performed under the the null hypothesis $H_0$ that no correlation exists between the quality of a resource and its usage, and the alternate hypothesis $H_a$ of the

---

3 The Mann-Whitney Wilcoxon test is a non-parametric test for differences between groups.
Table 6.4: Mapping of quality indicators of Bethard et al. [17] to signals that can be inferred before using a resource. An ‘X’ indicates that a signal for this indicator is not present.

<table>
<thead>
<tr>
<th>Quality indicator</th>
<th>Signal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has instructions</td>
<td>Description, Time tag</td>
</tr>
<tr>
<td>Identifies age range</td>
<td>Grade tag</td>
</tr>
<tr>
<td>Has prestigious sponsor</td>
<td>X</td>
</tr>
<tr>
<td>Has sponsor</td>
<td>X</td>
</tr>
<tr>
<td>Organized for learning goals</td>
<td>X</td>
</tr>
<tr>
<td>Content appropriate for age</td>
<td>X</td>
</tr>
<tr>
<td>Identifies learning goals</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 6.5: Signal weights

<table>
<thead>
<tr>
<th>Signal</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>60%</td>
</tr>
<tr>
<td>Time tag</td>
<td>20%</td>
</tr>
<tr>
<td>Grade tag</td>
<td>20%</td>
</tr>
</tbody>
</table>

true correlation being greater than 0 i.e. the higher the quality of a resource, the greater the number of users. My results show only a weak correlation of 0.124 between resource quality and usage ($t = 2.8343, df = 516, p = 0.002387$).

6.5.3 Social influence and resource usage

My next step explored whether social influence exerted by the decisions of others influenced the usage of resources. My first task in this regard was to determine if any correlation existed between the number of saves and the usage of a resource. A statistically significant positive correlation of 0.634 at a p-value of $2.2e^{-16}$ between saves and usage was found. This correlation is highlighted in Figure 6.6. Unlike earlier tests on resource position and quality, this provides preliminary evidence that the social influence conveyed through the saving of resources may be in part responsible for driving usage. This trend is even more apparent in Figure 6.7 which shows the number of saves and the average usage of resources over time.
6.5.3.1 Information cascade and resource usage

As described in section 2.3.4, an information cascade is a sequence of decisions where it is optimal for individuals to imitate the choice of others ahead of them. Bikhchandani, Hirshleifer and Welch (BHW) [18] introduced an information cascade model based on Bayesian reasoning which illustrates how information cascades can occur under rational conditions. The goal of this analysis is to determine if information cascades can generate a power law as observed in the usage distribution of community-contributed resources. To this end, I developed an information cascade model simulating the decision making process of educators. My model extends the BHW model in the following ways:

- Decision between multiple choices: Instead of the binary decision model of BHW, a decision will be made between 1..\textit{r} resources at any time
Figure 6.7: Trend lines showing the relationship between the number of unique users and saves on a resource

- Public signals: In the BHW model, the decision of an individual is always visible to others as a public signal. In the context of this research, the only public signal available is whether or not a user saves a resource. After clicking on a resource, users will leave public signals with a uniform random probability $p$. The value of this probability is exogenously fixed at 0.41. This was determined by computing the ratio of saves to unique clicks on all resources across all academic years.

- Private signals: A user’s private signal $p_s$ for a resource $r$ is drawn from a discrete uniform probability distribution such that $p_s \in [0, 1]$. If $p_s = 0$, then the agent infers a low private signal, otherwise if $p_s = 1$, the agent infers a high private signal.
Before describing the information cascade model I used, I begin with an illustration of the experiments of Anderson and Holt’s [10] to show how the BHW model can be used to understand how individuals can end up in an information cascade under rational conditions.

In the experiments of Anderson and Holt [10], a group of students sequentially decided if an urn containing 3 marbles was either majority-blue or majority-red. A majority-red urn contained 2 red marbles and 1 blue marble, and a majority-blue urn contained 2 blue marbles and 1 red marble. The urn was equally likely to be majority-red or majority-blue. After drawing a marble from the urn, students observed the color of the marble (private signal) and announced their decision i.e., whether the urn was majority-red or majority-blue to the rest of the class (public signal). If a student draws a blue marble, then she has a high ($H$) private signal for the majority-blue option and a low ($L$) private signal for the majority-red option. The converse occurs if she draws a red marble from the urn. An information cascade occurs when students ignored their private signals to follow the decisions of others before them. For example, if a student has a high private signal for the majority-blue urn—i.e. she drew a blue marble—but ignored this signal to declare the urn as majority red, then an information cascade is said to have occurred.

From the set-up of the experiment, we know the prior probabilities of a majority-red and majority-blue urn are each $1/2$. To ensure that all equations fit properly on the page, I shorten $majority - blue$ to $major - blue$ and $majority - red$ to $major - red$.

$$\Pr(major - blue) = \Pr(major - red) = 1/2 \quad (6.1)$$

Also, the posterior probabilities of choosing marbles that matches the actual composition of the urn are:

$$\Pr(blue|major - blue) = \Pr(red|major - red) = 2/3 \quad (6.2)$$
Posterior probabilities of choosing marbles that do not match the actual composition of the urn are:

\[ \Pr(\text{blue}|\text{major} - \text{red}) = \Pr(\text{red}|\text{major} - \text{blue}) = \frac{1}{3} \quad (6.3) \]

I now work through the sequential drawing process, to see how Bayes’s rule can be used to reason about the decisions of participants. Suppose the first student draws a red marble, she would want to determine if \( P(\text{majority} - \text{red}|\text{red}) > \frac{1}{2} \). Using Bayes rule this can be calculated as follows:

\[ \Pr(\text{major} - \text{red}|\text{red}) = \frac{\Pr(\text{major} - \text{red}) \cdot \Pr(\text{red}|\text{major} - \text{red})}{\Pr(\text{red})} \quad (6.4) \]

and

\[ \Pr(\text{red}) = \Pr(\text{red}|\text{major} - \text{red}) \cdot \Pr(\text{red}) + \Pr(\text{red}|\text{major} - \text{blue}) \cdot \Pr(\text{blue}) \]

\[ = \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{2} \quad (6.5) \]

If we plug this value in equation 6.4, we get:

\[ \Pr(\text{major} - \text{red}|\text{red}) = \frac{\frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2}} = \frac{2}{3} \]

Since this conditional probability is greater than \( \frac{1}{2} \), it makes sense for the student to decide that the urn is majority-red given a red observation. A similar calculation exists for the second student. The second will also announce a decision that matches her draw. If she draws a marble of the same color as the first student e.g., red, she announces that color as her decision. If she draws a different color, she now has two observations (i.e., the draw of the first student and hers), both of which are equally likely. Thus, we can assumes she flips a coin and goes with her observation. For the sake of brevity, I skip the mathematical details and suppose that second student announces a majority-red decision as well.

Now, I consider how the 3rd student might ignore her observation to follow the decisions of the previous two students. Suppose the third student draws a blue marble. Given
the prior two draws of red marbles, she will announce a majority-blue urn if
\[ \Pr(\text{major} \rightarrow \text{blue} | \text{red}, \text{red}, \text{blue}) > 1/2 \]
and a majority-red urn if
\[ \Pr(\text{major} \rightarrow \text{red} | \text{red}, \text{red}, \text{blue}) > 1/2. \]

Using Baye’s rule \( \Pr(\text{major} \rightarrow \text{blue} | \text{red}, \text{red}, \text{blue}) \) can be calculated as follows:
\[
\Pr(\text{major} \rightarrow \text{blue} | \text{red}, \text{red}, \text{blue}) = \frac{\Pr(\text{red}, \text{red}, \text{blue} | \text{major} \rightarrow \text{blue}), \Pr(\text{major} \rightarrow \text{blue})}{\Pr(\text{red}, \text{red}, \text{blue})} \tag{6.6}
\]

Since the draws from the urn are independent
\[
\Pr(\text{red}, \text{red}, \text{blue} | \text{major} \rightarrow \text{blue}) = \Pr(\text{red} | \text{major} \rightarrow \text{blue}) \times \Pr(\text{blue} | \text{major} \rightarrow \text{blue})
\]
\[
\Pr(\text{red}, \text{red}, \text{blue} | \text{major} \rightarrow \text{blue}) = \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{2}{27}
\]

\[
\Pr(\text{red}, \text{red}, \text{blue}) = \Pr(\text{major} \rightarrow \text{blue}) \times \Pr(\text{red}, \text{red}, \text{blue} | \text{major} \rightarrow \text{blue}) + \Pr(\text{major} \rightarrow \text{red}) \times \Pr(\text{red}, \text{red}, \text{blue} | \text{major} \rightarrow \text{red}) \tag{6.7}
\]
\[
\Pr(\text{red}, \text{red}, \text{blue}) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} + \frac{1}{2} \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{1}{9}
\]

Plugging this value back into Equation 6.6
\[
\Pr(\text{major} \rightarrow \text{blue} | \text{red}, \text{red}, \text{blue}) = \frac{2/27 \times 1/2}{1/9} = \frac{1}{3}
\]

Thus the third student will choose the majority-blue option with a \( \frac{1}{3} \) chance even though she observed a High signal for it. Since this value is less than \( \frac{1}{2} \), the student will instead choose the majority-red option disregarding her observation of a blue marble. This choice heralds the beginning of an information cascade.
6.5.3.2 Information cascade model

I now introduce my information cascade model with a description of the ingredients of the model:

- **Agents**: These represent the users (educators) in the observed dataset. An agent can click on and save a resource after clicking. I initialized the number of agents in the model to an arbitrary default of 154. This matches the number of educators that used community-contributed resources across all academic years.

- **Resources**: Resources can be clicked on and saved by agents. I initialize the number of resources in the model to an arbitrary default of 532. This matches the number of resources that were used across all academic years.

- **Private signals**: For each resource, an agent has a signal as to whether to use the resource or not. This signal is referred to as the agent’s private signal. Private signals can either be high ($H$) or low ($L$). An agent’s private signal for a particular resource is modeled as a random integer $N$ such that $0 \leq N \leq 1$. If $N = 0$, then the agent infers a low private signal, otherwise if $N = 1$, the agent infers a high private signal. Agents follow their private signal with a probability $s$. Thus, if an agent infers a high private signal for a particular resource, then with a probability $s$, the agent will click on the resource, and with a probability $1 - s$ an agent will not click on the resource. $s$ refers to the strength of an agent’s private signal. It is a real number between $0 - 1$. A value of 1 indicates a perfect private signal. Bikchandani et al. [18] and Anderson & Holt [10] assume that an agent’s private signal is perfect—i.e. has a probability of 1—by default. However, Bikchandani et al. [18] also experimented with private signal strengths ranging from 0.55 – 0.95, and they found that weaker private signal strengths took longer for an information cascade to converge on a particular option. In this experiment, I set the value of
an agent’s private signal strength $s$ to 1, i.e., agents will always follow their private signal.

- **Public signals**: Public signal of a resource is determined by the number of users who have saved it. Public signals allow users to infer about the decisions of others before clicking on a resource.

Agents decide on what resources to use (click on) and subsequently save in three steps. First, they select a resource from the set of all resources. Then, they make a decision as to whether to click on this resource. This decision is based on a combination of an agent’s private signal for a resource and the resource’s public signals (i.e., other agents that have saved the resource). Finally, if the agent decides to click on the resource, she leaves a public signal (saves it) with a $p$ of 0.41.

Unlike the BHW [18] model, a decision for a specific resource in my model has been split into two steps, which are: choosing a resource from the set of resources, and consequently deciding as to whether to click on it. In the binary decision process of the BHW model [18], a choice for one choice, by default, indicates a decision against the alternative. However, this decision making process is not amenable to situations where individuals choose between more than two alternatives. Thus, splitting the decision making process into two distinct processes where agents first select one out of $r$ resources, and then decide as to whether to use this resource provides a simple way of extending the decision making process of the BHW model [18].

The information cascade model evolves in $t$ discrete time steps or iterations. During each time step, the action of an agent occurs under one of two conditions:

**No public signal across all resources**

- **Selecting a resource**: The agent selects a resource at random from the set of resources with a uniform probability $1/r$ where $r$ is the total number of resources.
• **Clicking the selected resource**: The agent decides whether to click the selected resource according to her private signal. She clicks on the resource if she infers a high private signal. Otherwise, if she infers a low private signal, she does not click on it.

• **Saving the resource after clicking**: If an agent decides to click on a resource then with a probability of 0.41 she also saves the resource. This action leaves a public signal for subsequent agents.

**Public signals exist**: If at least one resource has been saved, then the decision process of agents is a bit different from when no public signals exist.

• **Selecting a resource**: The probability of selecting a resource is relative the number of public signals (saves) it has. Before any public signals were contributed to the set of resources, the probability of selecting any particular resource was a uniform probability $1/r$. After public signals exist, this probability is updated in favor of resources with public signals (saves). The greater the number of saves on a resource, the higher the likelihood of it being chosen.

• **Clicking the selected resource**: After selecting a resource, the agent evaluates whether to click on this resource against her private signal. If her private signal is high($H$), she proceeds to click on the resource. If her private signal is low($L$), she weighs the resource’s public signals (saves) against her private signal in deciding whether to click or not. If there is a single public signal (save) on the resource against her private signal of $L$, then the probability that she clicks on the resource is 0.5. In my model, I assume that in such cases, the agent sides with her private signal. On the other hand, if there are 3 public signals (saves) on the resource against her private signal of $L$, then the probability of click on the resource is $\frac{3}{4}$ and the the probability that the agent does not click on the resource is $\frac{1}{4}$. An information cascade occurs when an agent with a low private signal for a resource decides to click on it due to the preponderances of saves (public signals) the resource has.
• Saving the resource after clicking: After clicking on a resource, a save signal is left with a probability of 0.41

6.6 Results

2500 simulations of the information cascade model described above were processed with each simulation running for 2500 discrete time steps. Each of these simulations were evaluated to see if they follow a power law using the procedure of Clauset et al. [38] as described in section 6.4.2. As shown in Table 6.6, 82.6% of these simulations were determined to follow a power law distribution. The outcomes of this experiment strongly suggests that the above information cascade model of the decision making process of educators can lead to a power law distribution in the usage of resources. This finding provides credible evidence for social influence as a plausible hypothesis for the observed power law distribution in the usage of community-contributed resources.

Table 6.6: Results of information cascade simulation experiments

<table>
<thead>
<tr>
<th>Stat</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of cascade simulations</td>
<td>2500</td>
</tr>
<tr>
<td>Number of simulations without a plausible power law fit</td>
<td>230</td>
</tr>
<tr>
<td>(step 2 of Clauset et. al[38] procedure)</td>
<td></td>
</tr>
<tr>
<td>Number of simulations where alternative distributions</td>
<td>205</td>
</tr>
<tr>
<td>provided a better fit compared to the power law</td>
<td></td>
</tr>
<tr>
<td>(step 3 of Clauset et. al[38] procedure)</td>
<td></td>
</tr>
<tr>
<td>Number of good power law fits</td>
<td>2065</td>
</tr>
<tr>
<td>Simulations that follow a power law</td>
<td>82.6%</td>
</tr>
<tr>
<td>Simulations that do not follow a power law</td>
<td>17.4%</td>
</tr>
</tbody>
</table>

To ensure that the results of this model are highly unlikely by chance, I compared it to a random model. This model generated a set of 532 random integers with values ranging from 0-154. These 532 integers represented the 532 resources in the information cascade model and the value of these integers represented usage. Like the information cascade model, 2500 simulations of these numbers were created. Consequently, each of these simulations was tested to see if it follows a power law per the procedure of section 6.4.2.
None of the 2500 simulations of random integers fit a power law distribution. This further supports the hypothesis of information cascades as a plausible generative mechanism behind the observed power law distribution in the usage of community-contributed resources.

6.7 Discussion

In this study, I have demonstrated that an information cascade model simulating the decision making process of educators using community-contributed resources can generate a power law as observed in the empirical usage data. This strongly suggests that the hypothesis of social influence driving the usage of these resources is plausible.

This finding has important implications for both educators and agencies that support online learning communities.

For educators, it highlights the fact that popularity does not always equal quality. As indicated by multiple studies, the decisions of individuals acting in sequence can be inaccurate [18, 50]. Thus, educators should view a resource’s popularity as only a precautionary indicator of its value. Equivalent emphasis should also be placed on other pieces of metadata such as the description of the resource.

For agencies that support online learning communities, this research has important implications for resource presentation and recommendation. In presenting resources, social influence signals can be de-emphasized to limit the chances that they will detract users from evaluating a resource’s inherent quality. For example, in the CCS, the number of educators that have saved a resource can be hidden and require active effort from users to be revealed. In recommending resources, high quality but barely used resources can be recommended to educators in ways that give them high precedence. This could include personalized recommendations while active on the platform or email recommendations.
6.8 Limitations

In this section, I discuss the limitations of the analyses conducted as part of this study and their implications.

6.8.1 Size of the datasets

In comparison to contemporary studies on information cascades and diffusion (spread) in social networks [9, 83, 33], the size of the datasets analyzed in this research may appear small. This, however, is a direct function of the research context: patterns of sharing among educators in a single school district. While “big data” analyses may be useful in understanding grand diffusion patterns, these techniques may not generalize to smaller, more focused communities. For example, models for understanding the diffusion patterns of content shared on Facebook [33] may not be directly applicable to the diffusion of educational resources in a learning community. One reason why this may be the case is that there is a much higher threshold for adoption in a learning community compared to Facebook. In learning communities, educators are focused on finding high quality resources that can improve their instruction, while on Facebook information shared may primarily serve entertainment purposes.

6.8.2 Logged position of community-contributed resources

Understanding the impact of a resource’s position and its usage is complicated by the fact that the CCS’s click tracker only records the position of a resource in a list and not the page of the list in which the resource belongs. A list of resources is divided into pages with each page containing up to 10 resources (positions 0-9). Thus two items of the same position may actually be located on different pages. In light of this limitation, I made the assumption that all resources are located on a single page with positions 0-9. The impact of this limitation is that the rigor of findings from this study may not be applicable to the
usage of resources beyond the first page of the list they appear in. However, this limitation may be mitigated by findings from research on the browsing behavior of users online which indicates that 91% of all web searchers do not go past the first page of search results [127]. If we assume that educators in the CCS follow a similar behavior, then the outcomes of this study will be largely indicative of the usage behaviors of educators in the CCS.

6.8.3 Social influence at the individual level

In this study, as with prior work on information cascades [18, 10, 46], social influence was examined at the aggregate level. I demonstrated—through the lens of information cascades—that the decision of an educator to use a particular resource may be in part due to the aggregate decisions of prior educators to also use the same resource.

However, in reality, the choices of an individual’s close neighbors, family or friends may have a stronger influence on his decision as compared to the aggregate decision of everyone else. Consider the example of an information cascade that was introduced in section 2.3.4, where an individual $i$ abandons his preferred restaurant $A$ to follow the decisions of others and dine at restaurant $B$. Now imagine, that $i$ has a close friend $j$ whose preferred restaurant choice is also $A$. Unlike $i$, $j$ is not easily swayed by the decisions of others. Thus, even though $A$ is relatively empty as compared to $B$, $j$ insists that he and $i$ should still dine at $A$. Although $i$ has a penchant for following the crowd, he may decide to ignore this feeling and dine at $A$ since his friend $j$ insists that they both dine there.

Several prior studies have considered the effect of person-to-person influence on the decision making of individuals. Ryan and Gross study on the adoption of hybrid corn seeds among farmers in Iowa showed that while most farmers learned about the seeds from salesmen, they were only convinced to try out the seeds when their neighbors also used the seeds [50, 112]. One might imagine that the salesmen would probably had informed the farmers of how popular the hybrid corn seeds were. However, the farmers weren’t convinced of the value of the hybrid corn seeds until they observed a close neighbor also adopt the hybrid
corn seeds. Similarly, in a study of the adoption of tetracycline by physicians, Coleman et al. [50, 40] showed that the physicians were more likely to adopt tetracycline as a part of their treatment regimen for patients if their close colleagues also did.

In the context of the CCS, we have no way of understanding all the influences that play a role in an educator’s decision to use a resource. In particular, we do not know an educator’s ego network, in terms of how they decide on what resources to use in their instruction. We do not know who educators’ consult, and the weight of this influence on their decisions. Thus, the aggregate decision of others in the online professional learning community is our closest proxy for understanding the impact of social influence.

6.9 Conclusion

In this study, I evaluated the usage distribution of resources shared amongst educators in an online learning community. I sought to understand what this usage distribution was and what generative mechanisms may underly the observed distribution. Results indicate the usage distribution of resources followed a power law. Furthermore, I discovered that social influence investigated through the lens of information cascades provide a plausible generative mechanism for the observed distribution. 82.16% of 2500 simulations of an information cascade model simulating the decision making process of educators resulted in a power law distribution as observed in the usage distribution of community-contributed resources.
Chapter 7

Final thoughts

I began this dissertation with an outline of the three core components of my thesis. These are that:

(1) Sociological network theory can be used to explore the phenomena of homophily and triadic closures in online learning communities wherein ties between members are not evident.

(2) An understanding of the triadic closure process can be used to improve the recommendation of resources in these communities.

(3) Social influence may play a significant role in the diffusion and popularity of resources within online learning communities.

The outcomes of the three studies performed as part of this dissertation research provide strong evidence to support this thesis. I now summarize each of these studies and how it supports my thesis.

The first study of this dissertation (see chapter 4) showed that Granovetter’s theory on the strength of weak ties provides a useful theoretical lens for understanding the resource usage and sharing patterns of educators in an online learning community. Granovetter’s theory was operationalized on a deduced social network that was generated by creating edges between educators (nodes) based on their common usage of community-contributed
resources. The results of this study provided empirical evidence of the presence of homophily
and triadic closures in the community of educators that used the CCS.

Specifically, I showed that educators connected by stronger ties—ties with higher edge
weights—in the deduced social network shared greater similarity in their propensity to use
and share resources and level of comfort and use of technology in comparison to educators
connected by weaker ties—ties with lower edge weights. This confirms the presence of
homophily in the network, i.e., the notion that the strength of a tie between two educators is
directly related to the level of similarity between them. Furthermore, I showed that triadic
closures in the deduced social network are not random and are highly predictable. My results
indicate that the betweenness centrality, node degree and average edge weight of the common
neighbor shared by two unconnected nodes in a triad can be used to predict whether an edge
will form between them in the future.

The second study of this dissertation (see chapter 5) showed that an understanding
of the triadic closure process can be used to improve traditional resource recommendation
systems. Specifically, I showed that a traditional collaborative-filtering recommendation
system that is augmented with a triadic closure prediction model can provide statistically
significant increases in prediction accuracy. This improvement persisted even when the
collaborative-filtering system was combined with a content-based system to create a hybrid
recommendation system.

The third and final study of this dissertation (see chapter 6) investigated the usage
distribution of community-contributed resources and underlying mechanisms that may have
generated the observed distribution. The results of this study indicates that the usage
distribution of community-contributed resources is heavy-tailed, i.e. a few resources are
widely used while most resources were barely used. In particular, the usage distribution of
community-contributed resources follows a specific class of heavy-tailed distributions known
as power laws.
Investigations into the underlying mechanism(s) behind the observed power law distribution in the usage of community-contributed resources indicate that neither the position of a resource nor its perceived quality played a significant role in its usage. However, social influence—investigated through the lens of information cascades—was found to play a significant role in the observed power law distribution in the usage of resources. Educators were likely to use community-contributed resources that they know that their peers have also used.

7.0.1 Research contributions

The three studies conducted as part of this dissertation make intellectual contributions to the fields of social network analysis, recommendation systems and economic theory on information cascades respectively.

7.0.1.1 Intellectual contributions

My first study demonstrated how social network analysis techniques coupled with sociological network theory can be used to understand the phenomena of homophily and triadic closures in online learning communities where ties between members are not evident.

My second study showed that traditional educational recommender systems can be improved by the incorporation of a computational model for predicting triadic closures. It suggests that educators are also likely to be interested in resources that have been used by others with whom they have a common neighbor that they share a similar usage history with.

Finally, my third study contributed a computation model for understanding the usage behavior of educators that can potentially lead to inequity in the usage distribution of community-contributed resources. In particular, I extended the classic information cascade model of Bikchandani, Hirshleifer and Welch (BHW) [18] to a scenario where individuals select among multiple choices and when only an aggregate of the decisions of prior individuals
is available to them. This extensions confirm suggestions of Bikchandani et al. [18] that information cascades can still occur under this scenario.

7.0.1.2 Implications for school districts and agencies that support online learning communities

In addition to the aforementioned intellectual contributions, findings from this dissertation will be beneficial to educational agencies, such as school districts that are invested in online professional learning communities.

By coupling field research with usage log data, I showed that educators that participated in the CCS’s online learning community demonstrated a higher perceived level of isolation in comparison to educators that did not use community-contributed resources. This finding suggests that school districts may be able to mitigate the isolation felt by their educators by encouraging and facilitating their use of online learning communities. Many educators in the school district studied in this dissertation were the only Earth Science educator in their school. Thus, the CCS’s learning community provided a convenient way for them to share and learn from their peers. Multiple studies indicate that teacher isolation can have a detrimental impact on teacher enthusiasm, performance and ultimately student learning outcomes [34, 99].

Furthermore, an understanding of the factors that impact the usage of resources in online learning communities can be used to promote the usage of high-quality but barely used resources. As illustrated, social influence is a dominant factor in the usage of community-contributed resources. Like other domains, the impact of social influence can shroud user perception of unpopular yet high quality resources. In light of this, agencies that support online learning communities can incorporate alternative content delivery strategies to ensure greater equity in the usage of resources in their communities. These strategies can include personalized resource recommendations among others.
7.0.2 Potential future directions

I now discuss potential future directions to the research presented in this dissertation.

7.0.2.1 Do weak ties help alleviate educator perceived of isolation?

In the first study of this research (see Chapter 4), it was discovered that educators who used community-contributed resources, on average, had a higher perceived level of isolation in the classroom as compared to educators who did not use community-contributed resources. Unfortunately, the survey questions that were used to ascertain an educator’s perceived level of isolation were not asked in the post-survey interview that was conducted at the end of the 2011-2012 academic year.

Future work that can ascertain as to whether using community-contributed resources makes educators feel less isolated in the classroom will be very interesting and of added value to the work presented here. Prior research studies, have shown that in-person professional learning communities can help alleviate teacher’s perceived feelings of isolation in the classroom [34, 99]. However, there is a dearth of research suggesting if this effects persists within online learning communities.

7.0.2.2 Qualitative studies on the impact of social influence on educators’ usage of community-contributed resources

In the final study of this dissertation (see chapter 6), I showed that social influence may be a driving factor behind the usage of community-contributed resources, and hence the observed power law distribution in the usage of these resources.

While an information cascade model has illustrated that the social influence hypothesis is plausible, qualitative studies (such as interviews and surveys) that ask educators’ about their decision making process in using resources can be valuable as well. A preponderance of educators who indicate that the decisions of others play a major role in their decision on
what resources to click on will provide further validation for the social influence hypothesis.

7.0.2.3 Educator awareness of participation in an online professional learning community

In this dissertation, I have not explicitly considered the influence of educator awareness of the online professional learning community on participation. It is natural to wonder if educators equated participating in knowledge sharing with others in the CCS to participation in physical professional learning communities that they may have been a part of. If not, why so? Or it may be the case that educators felt that the CCS constituted a professional learning community but to a much lesser degree as compared to other professional learning communities they engage in. Prior research indicates that effective community of practices are characterized by individuals that are aware of their social presence, comfortable and motivated to share resources with others and are willing to collaborate with each other [30, 128].

While most educators who had access to the CCS were involved in training sessions on using the CCS to facilitate knowledge sharing with others in their district, the extent to which these training sessions influenced their perception of the CCS’s online professional learning community was not explicitly measured. Furthermore, there is no explicit reward for educators that participate in the CCS’s online learning community. For example, educators that contribute resources do not receive special recognition from the school district. Thus, external factors of motivation that may impact an educator’s participation were not considered in this research.

Future work that considers the role of educator awareness and motivations on their participation in an online professional learning community may shed greater insights into their usage behaviors and the community’s evolution over time. Research in other online communities such as Facebook has shown that behaviors of individuals within these communities is impacted by their awareness of the ramifications of their actions within the
community [2]. An early study of college students on Facebook noted that most students were unaware of the visibility of their actions (posts on Facebook) to others outside their immediate friend circle. Consequently, upon discovering the potentially wider audience of their Facebook posts, students that were concerned with privacy either tightened the privacy controls around their profiles or restricted the type of information they made public [2]. Perhaps, a similar phenomenon may can be observed in the CCS. For example, it may be the case that educators that are more vested in the online professional learning community as a core part of their instruction may be less swayed by the decisions of prior educators in choosing which resources to click on.

7.0.2.4 Richer representations of an agent’s private signal

In the third study of this dissertation, an educator’s private signal was modeled as a random integer $i \in [0, 1]$, where 0 indicated that an educator had a low private signal for a resource and 1 indicated that the educator had a high private signal for the resource. While straightforward, this simple duality of an educator’s preference (private signal) does not capture the innate complexity that make up the decision making process of educators. For example, it does not account for preferences of a particular resource type that educators may have. One can imagine that an educator who has a penchant for using animations and interactive visuals in her classroom may have a high private signal resource for resources of that nature as compared to word documents or power point presentations. Similarly, an educator’s need in the classroom at the time of using a resource can play a key role in her preference for a particular resource. An educator looking for something flashy and instantly engaging for her students will probably have a preference (high private signal) for an interactive resource such as an animation or video as apposed to a static resource such as a word document.

While capturing richer representations of an educator’s private signal may lead to improved information cascade models, such representations may be very difficult to obtain.
This is because many of the processes that impact an educator’s preference (private signal) for a resource are unobservable. For example, it is almost impossible to know an educator’s exact need at the time of using a resource. Furthermore, as with the cascade models of Bikchandani et al. [18] and Devany and Lee [46], casting the private signal of an individual as either high or low in the random fashion as done in this research, allows for information cascade models that are easy to understand and are generalizable to other domains.

7.0.2.5 Generalizability of findings

Finally, it will be interesting to explore the generalizability of the ideas presented in this dissertation to other domains. These could include other educational settings (e.g., a school district of Mathematics educators), or professional learning communities in institutions such as banks and engineering firms.

7.1 Acknowledgements

The research presented in this dissertation has been supported by National Science Foundation awards #1043638 and #1147590.
Bibliography


[34] Karin Chenoweth. It can be done, it’s being done, and here’s how. Phi Delta Kappan, 91(1):38–43, 2009.


[58] Adam Ginsburg. plfit 1.0.2: Python package index.


Appendix A

CCS pre-deployment survey questionnaire

This section outlines 37 questions that were asked to CCS’s participants as part of a pre-deployment survey questionnaire before the 2011-2012 school year. These questions were grouped under the following six categories:

(1) Class size
(2) Years teaching
(3) Class needs
(4) Level of comfort and use of technology when teaching
(5) Perceived level of isolation
(6) Propensity to use and share resources

Here are the questions that were asked under each of these categories

A.1 Years teaching

(1) The following question is for statistical purposes ONLY. Including the current school year, how many years of total teaching experience do you have?

(2) The following question is for statistical purposes ONLY. Including the current school year, how many years of total teaching experience in earth science do you have?
A.2 Class size

(1) What is your average earth science class size?

A.3 Level of comfort and use of technology when teaching

Rate your frequency of use of the following computer technologies (options: Rarely, A few times per semester, A few times per month, A few times per week, Daily)

(1) Microsoft Office (Word, Excel, PowerPoint) or similar programs

(2) An Internet search engine (e.g., Google, Yahoo!)

(3) A digital library (e.g., NSDL, DLESE)

(4) A favorite web site (e.g., NASA, NSTA, other)

(5) A social networking site (e.g., LinkedIn, Facebook, MySpace, other)

(6) A streaming video site (e.g., YouTube, TeacherTube) wife died and he wrote Paradise Regained.

(7) The Curriculum Customization Service (CCS)

(8) District or school made sites

Rate your level of comfort with using the following computer technologies in your instruction (options: Uncomfortable, Neutral, Comfortable, Very comfortable):

(1) Microsoft Office (Word, Excel, PowerPoint) or similar programs

(2) An Internet search engine (e.g., Google, Yahoo!)

(3) A digital library (e.g., NSDL, DLESE)

(4) A favorite web site (e.g., NASA, NSTA, other)
(5) A social networking site (e.g., LinkedIn, Facebook, MySpace, other)

(6) A streaming video site (e.g., YouTube, TeacherTube)

(7) The Curriculum Customization Service (CCS)

(8) Other (specify below)

A.4 Propensity to use and share resources

In their previous semester of teaching, participants were asked about to assign a frequency to the following questions (options: Rarely, A few times per semester, A few times per month, A few times per week, Daily)

(1) How often do you used materials created by other Earth Science educators in your district?

(2) How often do you look at materials created by other Earth Science educators in your district for inspiration?

(3) How often did you share materials that you created such as handouts, Powerpoint slides, rubrics, etc.-with other educators in your district?

(4) I have felt very comfortable sharing my materials and ideas with other Earth science educators in my district.

(5) It has been easy to share materials and ideas with other Earth science educators in my district.

(6) Do you agree with the following question: Sharing best practices and good ideas has been a routine practice among Earth science educators in my district. (options: Strong disagree, Disagree, Neutral, Agree, Strongly Agree)
A.5 Perceived level of isolation

Participants were asked to rate on a 5 point scale from Strongly Disagree to Strongly Agree on the following questions:

(1) I have opportunity to interact with other Earth Science teachers in my school

(2) I have opportunities to interact with other Earth Science teachers in my district

(3) I have opportunities to attend workshops and/or conferences

(4) I have a strong awareness of the curriculum practices of other Earth Science teachers in my district

(5) I have a strong awareness of the classroom instruction practices of other Earth Science teachers in my district

(6) I have a strong understanding of how other Earth Science educators in my district use interactive resources in their teaching.

A.6 Class needs

Participants were asked to rate on a 5 point scale from Strongly Disagree to Strongly Agree their agreement with the following questions when selecting materials for Earth science instruction. I need to consider the specific learning needs of these students:

(1) Individual or clusters of students with different knowledge, skills, or abilities

(2) Students with different reading abilities (e.g., ELA or Special Ed)

(3) Students with different quantitative skills

(4) Students with different cultural backgrounds and life experiences

(5) Gifted and Talented students
(6) Select which of the following describes your level of control in regard to curriculum of your classroom