André Miguel Guedelha Sabino

Master of Science

Potential Indirect Relationships in Productive Networks

Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy in
Computer Science

Adviser: Doutora Armanda Rodrigues,
Professora Auxiliar,
Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

December, 2016

Potential Indirect Relationships in Productive Networks

To Daniel

Acknowledgements

I would like to thank my supervisor, Professor Doctor Armanda Rodrigues, for all the advice, guidance, perseverance, and friendship, which were instrumental to the successful accomplishment of the work presented in this thesis.

A special thanks to Miguel Goulão, João Gouveia and João Rosa for their contribution to this work, and to all co-workers at the Interactive Multimedia Group, NOVA-LINCS, who made the office a friendly workplace: Nuno Correia, Carmen Morgado, Diogo Cabral, Rossana Santos, Rui Madeira, Filipa Peleja, Bruno Cardoso, Pedro Centieiro, Inês Rodolfo, Flávio Martins, André Mourão, and Serhiy Moskovchuk.

I would like to thank the members of the HIDRALERTA project, for their contributions and cooperation with my research on the emergency management domain: Juana Fortes, Teresa Reis, Rui Capitão, Pedro Poseiro, Pedro Lopes, Tiago Garcia, José Ferreira, and Pedro Raposeiro.

A very special thanks to my decade long co-worker and friend Rui Nóbrega, and to Ângelo Cardoso, Filipe Curado, and Ricardo Cebola for their friendship and support.

All these years of research were possible because of the unconditional support I received from my family, to whom I give the most heartfelt thank you. To my parents, Ana Sabino and Fernando Sabino, and brother, João Sabino, because your love has always been there. To my loving wife, Sara Gancho, for sharing this journey with me, giving it purpose and together making us a fantastic team. To our son, Daniel Sabino, who we welcomed this final year, for making our life so much brighter. And to Luna, whose feline perspective always helped a great deal. You all deserve credit for this accomplishment.

This work is partly funded by Fundação para a Ciência e Tecnologia, Ministério da Educação e Ciência, Portugal, doctoral grant SFRH/BD/47403/2008, research grant PEst-OE/EEI/UI 0527/2011, and research grant PTDC/AAC-AMB/120702/2010.

Abstract

Productive Networks, such as Social Networks Services, organize evidence about human behavior. This evidence is independent of the network content type, and may support the discovery of new relationships between users and content, or with other users. These *indirect relationships* are important for recommendation systems, and systems where potential relationships between users and content (e.g., locations) is relevant, such as with the emergency management domain, where the discovery of relationships between users and locations on productive networks may enable the identification of population density variations, increasing the accuracy of emergency alerts.

This thesis presents a Productive Networks model, which enables the development of a methodology for indirect relationships discovery, using the metadata on the network, and avoiding the computational cost of content analysis. We designed and conducted a set of experiments to evaluate our proposals. Our results are twofold: firstly, the productive network model is sufficiently robust to represent a wide range of networks; secondly, the indirect relationship discovery methodology successfully identifies relevant relationships between users and content. We also present applications of the model and methodology in several contexts.

Keywords: Productive Networks, Machine Learning, Spatial Information

Resumo

Redes Productivas, tais como Redes Sociais, organizam evidência sobre comportamento humano. Esta evidência é independente do tipo de conteúdo da rede, e podem permitir a descoberta de relações potenciais entre utilizadores e conteúdo, ou entre utilizadores. Estas *relações indirectas* são importantes para sistemas de recomendação, e para sistemas onde relações potenciais entre utilizadores e conteúdo (e.g., localizações) são relevantes, tal como no domínio da gestão de emergências, onde a descoberta de relações entre utilizadores e localizações pode permitir a identificação de variações de densidade populacional, melhorando a precisão do nível de alerta.

Esta tese apresenta um modelo de Redes Productivas, que permite o desenvolvimento de metodologias para a descoberta de relações indirectas, utilizando exclusivamente metadados, e evitando o custo computacional da análise de conteúdos. Foram desenhadas e executadas experiências para avaliar estas propostas. Obtivemos dois resultados principais: em primeiro lugar, o modelo de redes productivas apresenta robustez suficiente para representar um leque variado de redes; em segundo lugar, a metodologia de descoberta de relações indirectas é bem sucedida a identificar relações relevantes entre utilizadores e conteúdo. Também apresentamos aplicações reais do modelo e metodologia.

Palavras-chave: Redes Produtivas, Aprendizagem Automática, Informação Espacial

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Summary

The motivation for this thesis is the need to improve the discovery of potential relationships between people. To address this issue we present a research plan, starting with a social network services survey to support the definition of a productive network model, enabling relationship discovery methods. This chapter presents our research questions and enumerates the main contributions of our work.

1.1 Background

Our previous work focused on tools for emergency plan construction and evaluation [52, 54, 58]. The main case study was a damn break scenario, where the emergency event originates at one point, and over time evolves through a large area, with different emergency management concerns. Planning for emergency scenarios requires input from several types of expert, which generally are familiar with a particular area, and specialized on a set of spatial, social, or economic characteristics.

One of the conclusions of our efforts was that the identification of experts planners was difficult, mainly because in this domain people seem to focus on a particular region. We asserted that, in Portugal, emergency plan construction follows a bottom up strategy, where a national emergency plan is a compilation of several regional plans, which in turn are the result of a composition of municipal plans. The bottom plans, which actually define the actions in the field, are built locally, isolating experts, which in turn makes some expertise on one particular scenario difficult to find.

This assessment motivated a more general question: how may we improve *relationship discovery* between people with similar occupations or interests? Particularly, this

discovery should be motivated by a person's *production*, and not necessarily by an explicit intention to increase the number of social relationships. Our intuition is that we may relate two persons using relationships between their production, which is increasingly present on social network services.

Another interesting conclusion of our work with emergency management was that while risk assessment is parametrized with, among other variables, population density, this is approximated by census data. Assessing the current population density in an area is a difficult problem, and predicting variations in the future is even more difficult. However, early warning systems could benefit from social network based population density variations estimations. Our idea is to use the information on social network services to support tools for people-location relationship discovery which may enable population density variation forecasts.

1.2 Context

There are several online services for content sharing. In fact, the number of services is increasing, with some focusing on different contexts of the social life, and others on the nature of the content media. This process of specialization is an indicator of the diverse nature of both content and sharing context.

These services are usually referred to as *social network services*. By definition, a social network "*refers to the ways in which people are connected to one another and how these connections create and define human society on all levels: the individual, the group, and the institutional*" [12]. Currently, the focus of social network services varies from the support of human relationships to the sharing of content. While supporting a set of similar features, these services do differ. We identified three types of services: *Social Networks*, *Content Sharing Networks*, and *Content Indexing Networks*. Chapter 3 presents the individual definitions of all these networks, and their shared context.

In this thesis, we refer to the set of Social Networks, Content Sharing Networks, and Content Indexing Networks as *productive networks*, because we conduct a study and conclude that these networks share information concepts and organization (see chapter 3), which supports a generic productive network model (see chapter 4).

Productive networks enable users to annotate content using keywords, which produces classification and categorization systems with many potential applications, such as profiling users according to interests, enabling advertising and e-commerce customization, identifying popular topics, or even locations of interest.

All elements that are produced by users, such as content items, and annotation elements such as keywords and geographic locations are referred to, in this thesis, as *artifacts*.

Networks have a natural representation as a graph (see figure 1.1). In the context of productive networks, each node is a network element, i.e., user, item or keyword. Edges of this graph may represent relationships between users, or between users and artifacts. We are interested in the structure and visibility of these relationships.

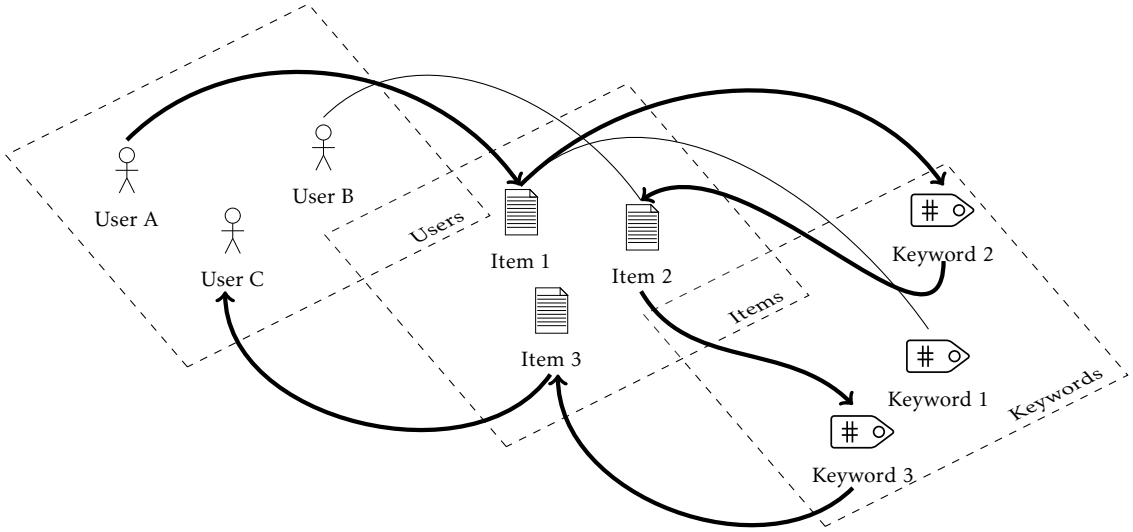

Figure 1.1: Graph of productive network elements. The bold path between users A and C includes different types of network elements.

The broad hypothesis of our work is that the *relationships between users and items, and the annotation processes present in productive networks constitute evidence of human behavior*. This evidence refers to the structure and organization of human relationships and content production, which is present in productive networks independently of the content media type.

Particularly, our focus is on *indirect relationships*, which we define as relationships that are not represented as a single edge or path of edges between the same type of elements in the network graph. These relationships are actually represented by a path involving all types of elements: from user to items; between items through annotation keywords; and back to users. Figure 1.1 illustrates a path between users A and C which is only possible through the relationships between the users' items and the respective annotation keywords. The bold path is indirect – it has to go back to the type of element.

In this thesis we systematically evaluate productive networks to define a theoretical model, which enables the construction of tools for the identification of potential indirect relationships between different network elements.

1.3 Problem Statement and Research Questions

The work in this thesis is focused on addressing the challenge of identifying indirect relationships on social networks, content sharing networks, and content indexing networks. Our goal is to show that these types of networks share a common data model, and present methods that enable the identification of indirect relationships from that common model. We formulate the main research question as:

Main Research Question: How to identify indirect relationships on social networks, content sharing networks, and content indexing networks?

We begin by identifying the common model between these three types of network through a *productive network* abstraction model. We formulate this research question as:

RQ 1. What are the common information concepts between social networks, content sharing networks and content indexing networks?

The answer to **RQ 1** constitutes the evidence to propose a network abstraction, the *productive network*, which represents the common aspects between the three types of network.

RQ 1 leads to the definition of a model suitable for supporting indirect relationships discovery:

RQ 2. How may we characterize productive networks in terms of the relationships between their underlying information concepts?

From the model definition, we focus on whether annotations associated with content items on productive networks enable the identification of relevant indirect relationships between users, or relevant indirect topics of interests for a user. More specifically:

RQ 3. May we identify relevant indirect topics of interest for productive network users?

RQ 4. May we identify potentially relevant indirect relationships between productive network users?

One of the focus while studying information concepts of productive networks (**RQ 1**) was that a subset of productive networks enables the association of geographic locations with users and content items. In this context, we ask:

RQ 5. May we specialize the model to account for location based annotation systems?

Focusing on potential applications of the model, we are interested in studying the effectiveness of retrieval tool based on the productive network model, designed to discover indirect artifacts:

RQ 6. What is the effectiveness of indirect artifact discovery methods based on the productive network model?

1.4 Research Overview

To address the challenges raised by the research questions, we designed a research plan that involves the definition of the productive network model supported by empiric observations of networks. The plan also included a series of studies which enabled the evaluation of the model as a supporting tool for relationship discovery. Ultimately, the plan guided our research to a preliminary stage of application as a population density variation estimation tool, specifically tailored for the emergency management domain.

1.4.1 Productive Network Survey

Driven by the possibility that there is a common structure between several types of networks, we designed a survey with the aim of capturing the data needed to gather the evidence required to propose a productive network model, which abstracts the type of network and represents that common structure. This survey partially answers **RQ 1**.

1.4.2 Formal Model Definition

Based on the evidence produced by the survey, we proceeded to define a model which should be able to represent all types of relationship between the concepts identified in the survey, i.e., users, content, annotations, and location references. Together with the productive network survey, the definition of the model answers **RQ 1**.

To further our understanding of the model we designed an interaction and visualization tool, which was evaluated by user trials. This tool, together with the complete model definition, answers **RQ 2** and **RQ 4**.

The evidence produced by the survey suggests that the spatial dimension is frequently available for content annotation. We extended the model to account for a new network element – locations – answering **RQ 5**.

1.4.3 Productive Network Sampling and Experiments

To answer **RQ 3**, **RQ 4**, and **RQ 5** we designed a set of experiments with real network samples. The experiments verify that the model enables the identification of relevant suggestions to the users. Furthermore, the experiments' outcome also enabled the evaluation of the effectiveness of the model as a supporting tool for information retrieval methods focused on indirect relationships discovery, answering **RQ 6**.

1.4.4 Applications to Emergency Management

The work related with **RQ5** and **RQ 6** enabled the discussion on the possibility of the transference of these results to real applications. We are interested in its impact on the emergency management domain, where the estimation of population density variations is key to determine the risk associated with emergency events. We will discuss how our

model enhances population density estimation, and outline a future integration with ongoing work in early warning systems.

1.5 Contributions

This thesis has the following contributions:

Productive network survey A study which surveys 41 networks, characterizing each according to a set of criteria relevant to describing its data model, relationship design, annotation strategy, and geo-referentiation strategy.

Productive network model A model which abstracts the network type and represents users, keywords and locations, and the relationships between them.

Indirect Relationships Discovery Methods Machine learning methods to discover indirect relationships, using indirect keywords or indirect locations.

Datasets Real network samples, in which the information extracted is represented according to the productive network model. These datasets were built to evaluate the indirect relationships discovery methods.

Contribution in other topics Research and development of an early warning system for coastal threats, which requires an assessment of the population density on the target area. This domain is a candidate for the application of several methods develop in this thesis, and we already achieved some preliminary results.

1.5.1 Peer-reviewed Publications

[53] Sabino A., Rodrigues, A., Productive Networks and Indirect Locations, in Leitner, M., Jokar Arsanjani, J., Citizen Empowered Mapping, Geotechnologies and the Environment, Springer Press, pending, 2017.

[50] J. Rosa, A. Sabino, and A. Rodrigues, Monitoring social network user density variations in areas of interest, in Proceedings of the 18th AGILE International Conference on Geographic Information Science, 2015.

[63] A. Sabino, J. Gouveia, and A. Rodrigues, Visualizing productive networks, International Journal on WWW/Internet, vol. 12, no. 2, pp. 34–50, 2015.

[20] J. Gouveia, A. Sabino, and A. Rodrigues, Visualizing productive networks relationships, in Proceedings of the 13th International Conference WWW/INTERNET, 2014

[57] A. Sabino and A. Rodrigues, Indirect Location Recommendation, in Proceedings of the 8th Workshop on Geographic Information Retrieval, 2014.

[59] Sabino, A., Rodrigues, A., Goulão, M., Gouveia, J., Indirect Keyword Relationships, in Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014.

[56] Sabino, A., Rodrigues, A., Understanding the role of cooperation in emergency plan construction, in Proceedings of the 8th International ISCRAM Conference, 2011

[55] Sabino, A., Space Aware Cooperative Environments, in Doctoral Consortium of the Collaboration Researchers International Working Group Conference, 2010.

1.5.1.1 Publications in the Emergency Management

[64] Sabino, A., Poseiro, P., Rodrigues, A., Reis, M. T., Fortes, C. J., Reis, R., and Araújo, J., Coastal Risk Forecast System, in Journal of Geographical Systems, pending, 2017.

[17] C. J. E. M. Fortes, M. T. Reis, P. Poseiro, J. A. Santos, T. Garcia, R. Capitão, L. Pinheiro, R. Reis, J. Craveiro, I. Lourenço, P. Lopes, A. Rodrigues, A. Sabino, J. C. Ferreira, S. Silva, P. Raposeiro, A. Simões, E. B. Azevedo, F. Vieira, M. D. C. Rodrigues, and C. P. Silva, Ferramenta de apoio à gestão costeira e portuária: o sistema hidralerta, in Proceedings of the VIII Congresso sobre Planeamento e Gestão das Zonas Costeiras dos Países de Expressão Portuguesa, 2015, pp. 1–18.

[62] A. Sabino, A. Rodrigues, P. Poseiro, M. T. Reis, C. J. Fortes, and R. Reis, Coastal Risk Forecast System, in Proceddings of the 1st International Conference on Geographic Information Systems Theory, Applications and Management, 2015.

[61] A. Sabino, A. Rodrigues, J. Araújo, P. Poseiro, M. T. Reis, and C. J. Fortes, Wave Overtopping Analysis and Early Warning Forecast System, in Proceedings of the conference on Computational Science and Its Applications – ICCSA 2014, 2014.

[15] C. J. E. M. Fortes, M. T. Reis, P. Poseiro, R. Capitão, J. A. Santos, L. V. Pinheiro, J. Craveiro, A. Rodrigues, A. Sabino, S. F. Silva, J. C. Ferreira, P. D. Raposeiro, C. Silva, M. C. Rodrigues, A. Simões, E. B. Azevedo, and F. Reis, HIDRALERTA Project – A Flood Forecast and Alert System in Coastal and Port Areas, in Proceddings of the IWA World Water Congress and Exhibition, 2014.

[48] P. Poseiro, A. Sabino, C. J. Fortes, M. T. Reis, and A. Rodrigues, Aplicação do sistema HIDRALERTA de previsão e alerta de inundações: Caso de estudo da Praia da Vitória, in Proceedings of the 12th Congresso da Água, 2014, no. 1.

[47] P. Poseiro, M. T. Reis, C. J. Fortes, A. Sabino, and A. Rodrigues, Aplicação do sistema HIDRALERTA de previsões e alerta de inundações: caso de estudo da Costa da Caparica, in Proceedings of the 3rd Jornadas de Engenharia Hidrográfica, 2014.

[16] C. J. E. M. Fortes, M. T. Reis, P. Poseiro, R. Capitão, J. A. Santos, L. P. Pinheiro, A. Rodrigues, A. Sabino, M. C. Rodrigues, P. D. Raposeiro, J. C. Ferreira, C.

Silva, A. Simões, and E. B. Azevedo, O Projeto HIDRALERTA – Sistema de previsão e alerta de inundações em zonas costeiras e portuárias, in Proceedings of the 8th Jornadas Portuguesas de Engenharia Costeira e Portuária, 2014.

[14] J. C. Ferreira, C. J. E. M. Fortes, M. T. Reis, P. Poseiro, A. Sabino, A. Rodrigues, S. F. Silva, J. A. Santos, R. Capitão, L. Pinheiro, J. Craveiro, P. D. Raposeiro, A. Simoes, E. B. Azevedo, M. C. Rodrigues, and C. Silva, Sistema de Previsão e Alerta de Inundações em Zonas Costeiras e Portuárias – O Projeto Hidralerta, in XVI Encontro da Rede de Estudos Ambientais em Países de Língua Portuguesa e III Seminário Internacional de Ciências do Ambiente e Sustentabilidade, 2014.

[6] C. Fortes, R. Reis, M. T. Reis, P. Poseiro, R. Capitão, L. Pinheiro, J. Craveiro, J. A. Santos, S. F. Silva, J. C. Ferreira, M. Martinho, A. Sabino, A. Rodrigues, P. Raposeiro, C. Silva, A. Simões, E. B. Azevedo, F. Vieira, and M. C. Rodrigues, Aplicação do Sistema HIDRALERTA na Avaliação do Risco Associado ao Galgamento no Porto da Praia da Vitória, in Actas do 3o Congresso Internacional de Riscos, 2014.

[49] M. T. Reis, P. Poseiro, C. J. E. M. Fortes, J. M. . Conde, E. Didier, A. Sabino, and A. Rodrigues, Risk Management in Maritime Structures, in Proceedings of the 8th International Conference on Management Science and Engineering Management, 2014.

1.6 Document Outline

This thesis is organized to address our research questions in order, from the data collection that enabled the model definition to its future applications.

Chapter 3 presents the productive network survey, which builds the body of evidence that supports our model. Chapter 4 presents the model's formal definition. Chapter 5 presents the indirect relationship discovery methodology, using the productive networks model and a machine learning framework. Chapter 6 presents several experiments designed to evaluate the model and methodology. Chapter 7 presents several applications using the productive network model, including our network sampling strategy, an interaction and visualization tool, and the preliminary work on population density estimation for emergency management applications.

The related work is presented in Chapter 2, and our final conclusions in Chapter 8.

Chapter 2

Literature Review

Summary

This chapter presents a literature review of the state-of-the-art relevant to this thesis, including discussions of our research related work. We also present the most relevant tools used by our methodology, namely the graph sampling method, and support vector machines. We conclude that the increasing offer of social networks, with support for keyword and spatio-temporal based annotation systems, providing semantic to the user's content, relationships, and activities, creates an opportunity to address questions about the user's potential interests, represented by either keywords or locations, as framed by our research questions.

Our research questions focus on three main research topics:

Social and collaborative network characterization: Studies the information elements and user dynamics supported by the networks, dealing with the challenges of modeling concepts and proposing theory for a very active and innovative industry. We look for context and influences which may help to identify the key elements required to support indirect relationship discovery.

Annotation and geo-referencing systems: Focus on the relationships between keyword and spatio-temporal annotations, and the retrieval requirements from social, content sharing and indexing networks. We are interested in annotations to represent relationships in the network, which depend on the semantics of the annotation process, particularly whether annotations represent explicit user interests, how the interests of different users may define relationships between them, and if they support a interest based relationship graph between users.

User, Content, keyword and location recommendations: Recommendation systems rely on the relationships between network elements, presenting models and methods to build ranked lists of recommendations that meet information needs. We review contributions presenting methods which retrieve elements from the network to build recommendations, from media content similarity analysis to machine learning methods, including hybrid approaches, and evaluate their potential for generalization and application to productive networks.

Several publications on these research topics address issues related with the following key concepts:

Information Retrieval: General categorization term for methods which address information needs from a collection of resources.

Collaborative Filtering: Methods which use information collected from many users to answer a query from one user.

Machine Learning: Computational methods which learn and predict information elements on a particular domain.

We use the following notation to represent network aspects addressed by several authors in this review:

$$
\begin{aligned}
U &= \{U_1, \cdots, U_t\} \textit{ is a set of users, } t \geq 1 \\
I &= \{I_1, \cdots, I_m\} \textit{ is a the set of items, } m \geq 1 \\
K &= \{K_1, \cdots, K_o\} \textit{ is a the set of possible keywords, } o \geq 1 \\
A &\subseteq U \times K \times I \textit{ is a set of annotations.}
\end{aligned}
$$

Most authors evaluate their methods using datasets from the following online services:

Del.icio.us Social bookmarking service, available at `del.icio.us`

Enron email dataset Email collection with contacts between senior managers at Enron, available `www.cs.cmu.edu/~./enron`

Facebook Social network, available at `www.facebook.com`

Flickr Social network focused on photo sharing, with support for annotations and georeferences, available at `www.flickr.com`

Foursquare Location based social network, with reviews and ratings of several points of interest, available at `foursquare.com`

LinkedIn Social network focused on professional relationships between users and organizations, available at `www.linkedin.com`

Meetup Location based social network focused on event organization and promotion, available at www.meetup.com

Picasa Social network focused on image shared and annotation, including geographic referencing, available at picasa.google.com

Reddit Social network focused on communities centered on common topics of interest, available at www.reddit.com

Sina Weibo Social network, available at www.weibo.com

Whrrl Former location based social network, available at whrrl.com until 2007

Twitter Social network, available at twitter.com

Wikipedia Collaboration supported online encyclopedia, available at www.wikipedia.org

Yelp Location based social network, with reviews and ratings of several businesses, available at www.yelp.com

When possible, we discuss evaluation results. Generally, these are presented in absolute values or charts. In the publications focused on information retrieval problems, concerning the accuracy of queries, the most common metrics are the following (see chapter 6 for more detail on these and other metrics):

Precision The proportion of the returned results that are relevant;

Recall The proportion of relevant results that are returned;

F1 The harmonic mean of precision and recall.

We followed a systematic literature review protocol, whose parameters and quantitative analysis are presented in appendix D.

2.1 Network Characterization

Smith et al. [69] focus on social network characterization and dataset construction, in the context of enterprise organization dynamics analysis. Social network analysis may help measure the impact of corporate events, and human resource management in general. The authors address the state of social media proliferation in 2009, forecasting a widespread adoption in the professional context and, interestingly, an increase of location based or location aware networks.

In a study of existing systems and network design trends, the authors analyze several networks structures, proposing a set of dimensions with which network may the

characterized, which influenced our study (see chapter 3), namely which user roles exist, whether it supports hierarchies, and how it represents groups of users.

Gomez and Rogati [19] discuss how social network services are usually presented as tools changing how people interact with each other, and gain awareness of social events.

The authors question how real world events influence the evolution and temporal dynamics of networks, and present a study of how 10 thousand events influence a LinkedIn sample of 115 million users. All events are associated with a specific location, and because they are related with professional activities, there is a main topic of interest shared by the attending users. Results show that there is an increase in relationships immediately prior to the event. These effects should be taken into account for link prediction methods, and hint towards the use of location based network analysis for population density variation estimations in these cases. Section 7.2 presents our insight on this topic.

Atif et al. [41] and Viswanath et al. [72] present studies on user activity graphs evolution, and compare to relationship graphs. Activity graph present a different perspective on social network dynamics, which focus on the user production patterns over time.

One study analyzes 250 applications on Facebook [41]. Results show that activity graphs are directed, and tend to vary significantly over time. Ultimately, the authors propose a graph sampling method, and graph synthesis model which preserves the characteristics of activity graphs. This work, together with the work by Leskovec et al. [31] (see section 2.4.1) provides context for our sampling method.

Wilson et al. [78] in the context of social network study and analysis, present the question of how valid are social graphs to infer user interaction. Traditional social graphs represent explicit relationships between users, which will not guarantee the coverage of all social interactions between users. The authors present a study that constructs and analyzes a relationship dataset of 10 million Facebook users, which is enhanced with user behavior statistics, and interaction records. Conclusions show that links present in the social graph are significantly skewed towards a subset of the actual social interactions.

Burain et al. [5] present work on role identification. Users contribute content to networks in distinct contexts, and potentially assume different roles in different contexts. Identifying a user role in a context provides high level information about the user's topics of interest, and enables activity prediction methods. From a dataset of 279 Reddit users, the authors report that they in fact assume contextual roles, and tend to restrict their activity to one context.

Cruz et al. [10] address the problem of community detection in social networks. The authors present a model for community detection and visualization in social networks, in the framework by Wasserman and Faust [75], which describes networks as a composition of three types of variables:

Structural variable ($\mathcal{S}$): Information related with connections between users, with $\mathcal{S} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ represent users connected by relationships in $\mathcal{E}$;

Composition variable ($\mathscr{C}$): Information describing users, with $\mathscr{C} \in \mathbb{R}^f$, a vector of features;

Affiliation variable ($\mathscr{A}$): Information relating users to groups, with $\mathscr{A} = \{A_1, A_2, ..., A_k\}$, a collection of disjoint subsets of users.

A social network is formalized as $SN = (\mathscr{S}, \mathscr{C}, \mathscr{A})$. The model integrates this variables to enable data exploration and visualization tools from different perspectives. This model supports the notion of subset construction and different perspectives for network visualization, which is reflected in our visualization framework (see section 7.1).

2.1.1 Discussion

Social networks services are very popular in the present, and keep evolving and specializing, while keeping a common set of elements to represent users, their content, and annotation semantics [69]. Generically, users provide enough information to represent their interests and relationships [19].

While usually specialized in some activity domain or subject, these services represent the user activity and interactions [5, 41, 72, 78], and event community affinity [10], which enables tools to support and recommend further interactions.

2.2 Annotation and Georeferencing Systems

Heckner et al. [24] and Nov et al. [43] present work on how keywords are used and the motivation for their adoption. Heckner et al. conclude that the motivations for annotating are very specific of the network's specialization, for instance, items on publication indexing networks are often annotated with categorization keywords, while items in photo sharing networks are annotated with descriptive keywords. Nov, et al., report that annotating is actually a means to achieve social presence, mainly because keywords increase the chances of the user's content being discovered by others.

Hecker, et al., concluded that annotation behavior differs between systems, while also identifying annotation trends, such as that photographs are usually annotated regarding both content and location, and that both video and photographs are usually extensively annotated [24]. The authors single out Flickr users as the most adverse to the annotation process, while Nov, et al., concluded that Flickr users tend to use both descriptive and categorical annotations when their content audience is the general public, i.e., when the photographs are shared publicly.

Wu and Zhou [79] present a study on annotation systems, focused on the structure of the user network enabled by annotations. The authors study network properties, patterns of user behavior during the annotation process, the evolution of the actual annotations over time, and if and how the annotation supported user network reflects or influences user interests

To answer these research questions, the authors present a study of the Del.icio.us social bookmarking network. The working dataset contains 88 781 annotations, by an undisclosed number of users – generated from a seed of 275 users using 1 955 unique annotations. The dataset was analyzed from a network structure point of view, characterizing keyword relationships, and from a user point of view, describing behavior patterns. Results confirm several expected aspects of annotation systems, such as semantic proximity between connected keywords, and interest alignment between users who share keywords.

The dataset construction method is described as breadth first, which covered a high percentage of keywords of the network. This approach may be improved by crawling the network with the goal of preserving the network structure instead of keyword coverage. See section 2.4.1 for an alternative approach which preserves the original network structure in the sample.

Vijay and Jacob [71] present work on automatic image annotation, focusing on semantic gap mitigation through collaborative filtering. The authors review several content based methods, and all exhibit the same shortcoming: a systematic semantic gap between features extracted automatically, and the users manual annotations.

A candidate solution to the problem is a hybrid approach to automatic annotation, which combines content based feature extraction with user annotations [71]. The authors propose a hybrid approach, which includes feature extraction, content based retrieval of similar images, and keyword correlation probability estimation, which requires the computation of a keyword correlation matrix.

Sergieh et al. [66] present an hybrid method for geographic annotation of images. In this context there are two types of relationship between images: through annotations, and geographic distance. Two images may be annotated with the same set of keywords, but represent distinct geographical areas, or represent the same location and not share keywords.

This is a particular case of the automatic annotation problem, which also has to deal with the semantic gap problem. The authors propose to extract low level features from an image, and find similar images from which to recommend annotations.

Let I_t be an image of the location L_l, and $C = \{I_1, I_2, ..., I_m\}$ a set of images of the same geographical area, such that $\forall I_u \in C$, $geographical\_distance(I_t, I_u) < d$, where d is a predefined threshold. Also, members of C are annotated with keywords from $K = \{K_1, K_2, ..., K_o\}$. Therefore, the relationships between images and keywords are described by $\mathcal{G} = \langle C \cap W, \{(K_p, I_u) : K_p \in K \wedge I_u \in C\}\rangle$.

Extracting features from an image enables edges between the image and similar images. Therefore, a recommendation is supported by a path in $\mathcal{G}$ between the image and a keyword. The authors report precision values around 0.5 (reported in a chart) for 100 images, with 5 recommended keywords [66].

Hedge et al. [25] present a methodology to annotate locations with keywords. The

authors propose to use event data from location based social networks to extract information about locations using an unsupervised document clustering technique – the Latent Dirichlet Allocation.

The method is evaluated with information from two different types of network, i.e., event and location based networks, Meetup and Foursquare, respectively. Event descriptions on Meetup are processed to infer annotations, which are validated with the annotations on Foursquare, for a given location. Results report 67% of cases with relevant semantic relatedness [25].

2.2.1 Taxonomies

There is extensive work on keywords, their categories and usage patterns. Different systems enable different types of keyword organization and use permissions, from completely free to use and create approaches, to highly categorized and strictly curated environments.

Chi et al. [8] analyze the social annotation service Del.icio.us to show that, in that system, annotating is primarily a method of personal organization, with individuals using a personal vocabulary while annotating. The authors state that, although there is a clear lack of structure in that and in similar social annotation services, hints of a global language emerge at some point in those networks.

Zubiaga et al. [86] use information retrieval benchmarks to show that users whose keywords classify items outperform users whose keywords describe the content, which is compatible with our results, where a large number of keywords are associated with only one item, not serving as a classification system, but only as a descriptive one.

2.2.1.1 Folksonomies

Annotations are very much related with the term folksonomy (folk, or user, generated taxonomy). Thomas Vander Wall coined this term in 2004 [1], and is used to describe the user-generated taxonomies that became a distinctive aspect of Web 2.0.

To address the problem of lack of a clear and general semantic in folksonomies, Damme et al. [11] propose the study of *folksontology*. The authors discuss the main shortcoming of annotations as a classification tool, which is the lack of formal attachment between keywords and conceptual meanings or relationships. To overcome this issue, the authors propose the construction of ontologies from the social interaction manifested in folksonomies.

Xu et al. [81] discuss the use of keywords for the semantic web, deriving semantic correlations between related keywords on Flickr, which enables automatic image annotations. The authors propose to explore semantic correlations between keywords on a Flickr sample, and use a conditional random fields model to unify all the available information in an unified framework, which ultimately serves as an automated image annotation tool.

[1] http://vanderwal.net/folksonomy.html

Ilhan and Öğüdücü [39] assert that item recommendation generally represents an harder problem than keyword recommendation, but which makes sense to address for particular cases. The authors argue that for social bookmarking websites such as Del.icio.us, items may be completely represented by the collective annotation effort of the users. These annotations are unregulated, and their organization is best described as a folksonomy.

The authors argue that, although folksonomies yield problems such as ambiguities and redundancy, they create an opportunity to identify like-minded groups of people. They propose a recommendation engine were items are related with users through their relationships with other users and their keyword preferences. This approach relies on the identification of clusters of users, and the calculation of a user-item pair score.

Ilhan and Öğüdücü formulate a folksonomy as $F = (U, I, K, A)$, and propose to build bipartite graph clusters, presented as $S = (\mathcal{V}_a, \mathcal{V}_b, W, \psi)$, where $\mathcal{V}_x$ are vertex sets clusters such that a pre-determined clustering criteria function, ψ, is optimized, and W is the weight matrix with nonnegative weights corresponding to the edge between vertexes of $\mathcal{V}_a$ and $\mathcal{V}_b$.

The engine was evaluated with a collection of web pages annotated by on thousand users of the Del.icio.us social bookmarking site. Recommendation were validated using their cosine similarity with the users' annotations. Precision results vary from 0.4 to 0.5 in the report (reported in a chart, with a varying number of clusters). In our work, we follow a similar approach the keyword characterization problem for our raking requirements (see section 5.2), however the authors evaluation method is not comparable with ours, because it is based on a user trials, instead of a ground truth corpus.

2.2.2 Discussion

Annotation systems support the semantic annotation of content with keywords [8, 24, 43, 79], and some with geographical information [25, 66]. These keywords may be freely created by users [11, 39, 81], or through a curated process [8, 86].

Several authors conclude that there is an interest alignment between users who share keywords [79], and there is a semantic similarity or proximity between keywords annotating the same items [39]. Keywords may describe or categorize items, whether subject to a strict taxonomies, or as part of a user supported folksonomy. Ultimately, keywords represent the personal interest of the user, and enable connections between users.

2.3 Recommendation

We are interested in user, content, annotation and user recommendation methods, which are active research subjects. For example, Roth et al. [51] discuss a method to suggest contacts based on email contact lists and frequently used email addresses. Stefanidis et al. [70] discuss the use of preference contexts for group recommendation systems.

Other examples are [18, 33, 67], where several authors present approaches for the design of keyword recommendation systems. The authors focus on strategies to recommend keywords to enhance items' descriptions, and usually those keywords already annotate items of the user's contacts. Lappas et al. [29] discuss how social endorsement techniques can be used for keyword recommendation and ranking.

We are particularly focused on the relationships between the recommendation systems and the underlying network structure. Our goal is to identify the common requirements and assumptions for user, content and annotation recommendations.

2.3.1 User Recommendation

Chao Zhou et al. [84] discuss the problem of recommending users with similar interests, on social network services. The authors propose *UserRec*, a framework which models user interest, enabling interest-based user recommendation. *UserRec* relies on keyword graph based community detection, which is combined with user role analysis. Finally, the approximation of a user's interest distribution by another user is evaluated through a similarity measure, namely the Kullback-Leibler divergence.

UserRec is evaluated with a sample of the delicious [2] bookmarking service. We directly compare our results with this approach [84].

Centintas et al. [7] present a method for recommendations on professional social networks. The main challenges are the definition of a user profile, and proximity measurement between profiles. The authors propose a probabilistic model for latent content and graph class identification, based on a set of features describing the user in the network.

Zhang et al. [83] present work on group recommendation for location based social networks. Specifically, the authors ask how to recommend groups related with real world events. To address the problem, the authors present *PTARMIGAN*, a method which combines the social and geographic characterization of groups with the representation of past interactions between users and groups. *PTARMIGAN* uses latent factor model based matrix factorization to assign a rank to each group in a geographic area of interest (e.g., a city) for every user. The method is evaluated using a Meetup dataset on New York City (NYC) and Los Angeles (LA), USA, with 5 001 and 10 944 users, respectively. Best reported precision values are 0.15 for NYC and 0.16 for LA.

Li et al. [32] focus on user relationships prediction. The authors propose to determine the proximity between users on social networks.

For a user U, let $Y = [Y_1, Y_2, ..., Y_{|V|}]^T$ be the initial relationship proximities between U and a set of nodes in the network. The problem is formalized by the optimization of

$$arg\ min_{f \in \mathbb{R}^{|V|}} \{\Omega(f) + \mu \|f - y\|^2\}$$

where $f = [f_1, f_2, ..., f_{|V|}]^T$ are the new values of relationship proximities between the user and the set of nodes, with $\Omega(f)$ representing the the cost of the changes represented

[2] https://delicious.com/

in f, and $\mu > 0$ is a trade-off between f and Y. The method is evaluated with a dataset of 3 910 Sina Weibo users, with good performances (precision is reported in a non-standard scale, but the method outperforms most state of the art approaches). However, the optimization process must be computed for every change in relationship proximities, rendering this a very computationally expensive approach [32].

A particular case of user recommendation is community recommendation, which frequently requires method for community detection. Sachan et al. [65] present work on this context. From the relationships between users in a social graph, and the common keywords between them, the authors build a Bayesian model for latent community detection. Communities are modeled as distributions in the user interaction space, which requires extensive offline computations. The method is successfully evaluated with the same Enron email dataset used by [40].

Bakillah et al. [2] also present work on community detection, focusing on Twitter. Particularly, the authors compare how different definitions of the interaction graph influence the performance of the community detection method. The authors identify several types of interaction edges between users, namely, mentions, follow relations, and generic post content.

2.3.2 Content Recommendation

Lappas and Gunopulos [28] address the problem of recommending content on social endorsement networks, while proving explanations for those recommendations. The authors define social endorsement networks as systems where users assign their preference to content.

The authors present a method to provide explainable recommendations, which, for a a user submitted a query, retrieves the most popular group of items previously annotated with the keywords in the query. These groups are precomputed, where the set keywords is extracted from the network. Recommendations are delivered to users alongside the most popular keywords describing the group. These keywords represent an explanation to the recommendations.

The method is evaluation with datasets of 500 Twitter users, and 456 764 BLP authors, mainly focusing on the success of the group definition strategy, without reporting retrieval metrics [28]. Explaining recommendations is an interesting challenge, and we address the issue with our productive network visualization framework (see section 7.1).

Kefalas et al. [26] report that recently there have been several developments on location based social network content and user recommendation, which could benefit from systematic survey and categorization efforts. The authors have surveyed 16 networks and 43 recommendation algorithms, which where published or presented since 2010. Networks are categorized through several dimensions: the distribution platform; the personalization type (generic interface and/or personalized); system features, such as check-in tools, availability of map visualizations, and content duplicate detection; the

recommendation type, such as users, locations, keywords, and events.

The recommendation algorithms survey categorizes works according to personalization type, recommendation type, data factors/features, methodology and models and data representation. The main goal is to determine the recommendation strategy tendency in location based networks. Most algorithms in the survey focus on location recommendation, which may be included in a route recommendation goal, with only a few recommending activities/events alongside locations. Collaborative filtering is the most popular strategy. However, the authors do not actually compare algorithm performances because they lack an experimental protocol, and all algorithms wee published with evaluations on different datasets.

This survey aims at a systematic categorization of location based networks, and specific their recommendation strategies [26], and influenced our productive network study (see chapter 3), which presents networks using this and more dimensions, which we consider useful for our systematic network categorization goal.

2.3.2.1 Keyword recommendation

Liu et al. [34] discuss keyword ranking using a probabilistic approach, using Flickr as a case study. The authors propose to estimate relevance scores for each keyword of an image which, combined with a keyword similarity graph analysis, enables the construction of ranked lists of keywords. Wang et al. [74] improve the results of Liu, et al. [34]. The authors present a machine learning approach for keyword ranking, proposing a semi-supervised learning framework, based on linear regression models, to rank annotation keywords of an image.

Mesnage and Carman [38] present a Bayesian model for keyword proximity. The main goal is to deliver keyword clouds to social network users. The problem is presented as the probability that a keyword, K, is relevant for a given query, Q, such that

$$P(K|Q) = \frac{P(Q|K)P(K)}{P(Q)}$$

where $P(K)$ corresponds to the global frequency of K, while $P(Q|K)$ translates into the probability that K is relevant given that the keywords in Q are relevant.

The authors propose three methods to compute $P(Q|K)$. The first computes the co-occurrence of K in the network. The second computes the frequency K is used by the user's relationships. The last method extracts latent topics from the set of keywords related with each user, and compute the probability that K belongs to a topic in the query. The methods are evaluated by users trials on a prototype application [38].

Zhou et al. [42] address the problem of automatic photo annotations on large, user contributed datasets. The authors starting point is the traditional content based image retrieval approach, where features are extracted from the media and then used to infer semantic annotations. This approach suffers from the previously mentioned semantic gap problem, and high computational costs.

The authors explore a possible solution for the semantic gap problem, by introducing user generated annotations. The approach is called *hybrid probabilistic model for automatic [keyword] recommendation*, and the idea is to integrate the low level information with the user generated annotations. This integration explores the correlations between low level based annotations and user annotations, to recommend keywords to images without user annotations.

Given that each image I_j is labeled with K_{I_j} keywords, $\{K_{I_j}1, K_{I_j}2..., K_{I_j}\}$, the keyword records may be represented as an $m \times o$ association matrix. The goal is to discover all situations where an association between a keyword and an image should be present in the matrix, but it is missing.

The hybrid probabilistic model calculates the posterior probability of a keyword being assigned to an image, based on its correlation with the the current set of keywords of the image. If the image has no keywords, it scales back to low level feature analysis.

The approach is evaluated with three datasets: two from the Corel collection, with 5 000 and 30 000 images, and a one built from Flikr, with 269 648 images. The Corel 5 000 image dataset recorded a best average precision of 0.47, with an average recall of 0.33. The Corel 30 000 image dataset recorded an average precision of 0.19 and an average recall of 0.38. The results of the Flikr dataset were not reported [42]. Although we do not focus on low level feature extraction, we formulate the problem of keyword recommendation (as an instance of network element relationship discovery) in networks using a similar approach as this work, and evaluate our methods with comparable methods. See section 6.2.3 for a performance comparison.

Xu et al. [80] present a method to increase the retrieval performance on annotation systems, relating keywords chosen by different users to annotate the same content. The method builds a K nearest neighborhood graph of keywords, estimating the kernel density of each keyword in its neighborhood. Keywords are scored by its kernel density, and the density of its connected keywords, which ultimately enables the identification of a most relevant keyword to represent the cluster. These keywords are then used in a depth first search to include all cluster members.

2.3.2.2 Location Recommendation

Laere et al. [27] use machine learning methods to automatically assign geographic coordinates to Flickr photos. The authors also use a clustering approach to obtain regions of interest, and present a method that successfully predicts the location of a previously unseen photo. The authors also provide a discussion on the effects of spatial granularity on the meaning of the location recommendation for a particular photo.

Ye et al. [82] present a method for semantic annotation for location based social networks. The method builds support vector machines for each keyword in the network, after an extensive feature extraction computation, combining user check-in behavior with inter location relationships. The method is evaluated using a dataset with 5 892 users

of the discontinued Whrrl social network. The evaluation is performed for categories of keywords (from Yelp, another network), instead of single keywords, reporting an average precision around 0.8 (reported in a chart).

Wan et al. [73] propose the integration of geographic information with keyword annotations to improve item retrieval. The authors propose that the geographic information could help distinguish between different semantics of the keyword.

For each keyword, the method would partition the geographic space such that the keyword has the same meaning in each partition. This property is asserted through the use of the keyword by users in close proximity, and the co-occurrence with other keywords, i.e., for keyword K_a, with co-annotates an item with keyword K_b, the locations associated with items also annotated with keyword K_b should be in the same region. The authors report a very well performance on a dataset with 8 119 Flickr users, and another with 509 Picasa users, with precision at rank 1 around 0.85. The evaluation does not use a ground truth corpus, but a user trial instead (10 users evaluated 10 queries each) [73].

Peregrino et al. [45] present a method to infer location from Twitter posts. It is based on text analysis, and cross referencing with Wikipedia entries. Our model and this work can be integrated in a solution that first infers the geographic location of a Twitter post, and then discovers related indirect locations for recommendation.

Ozdikis et al. [44] use evidential reasoning techniques over Twitter data to estimate locations, also with the ultimate goal of event detection. Using the Dempster-Shafer Theory, the authors use Twitter post locations, text, and user profile declared locations to construct belief intervals for sets of locations where certain events might have happened. This approach enables the discovery of locations that may be relevant to a user interested in a particular event. The evaluation is presented using belief percentage as effectiveness metric, which cannot be compared with our results.

Hauff [23] addresses the problem of retrieving items with geographic annotations, which are highly influenced by the quality of annotations, i.e., the accuracy of the latitude/longitude assigned to items. The author present a study on an image and video dataset from Flickr. The study collected images related with reasonably popular locations, and distinctively indoor contexts (to easy the annotation process). It is also restricted to limited sized venues, excluding vast natural landscapes, which would prove difficult to measure accuracy of coordinates. The ground truth used is the Wikipedia coordinates stated in the venue entry. Conclusions show that most geographic annotations are fairly accurate, with a significant increase for indoor pictures of popular locations. For popular locations, the average distance to the actual coordinates is 13 meters, and for the less popular is 167 meters, which hint at the limits of the precision assumptions of applications relying on such data sources. In our application prototypes, we keep our search queries limited to 1 km radius search areas or more (see section 7.2.1).

Wei et al. [76] improve on point of interest recommendation for location based social networks. Instead of the conventional approach of establishing a relationship between user and location, the authors propose to describe how groups of users relate to locations,

enabling a new type of dimension to support recommendations. The approach is evaluated in terms of how close together are the locations related with the same group. There is no report of recommendation metrics to compare with our results.

Phan et al. [46] also focus on the location recommendation problem, with a collaborative filtering approach. Specifically, the authors ask how to locations of interest to users, based on information from a set of previously annotated items by a large set of users.

Given a $U \times L$ co-occurrence matrix, the authors calculate the correlation between the users items geographic annotations with other locations used to annotate similar items. In this context, two items are similar if they annotate with the same locations. The actual recommendation system takes into account the user's current location, and filters the list with a given radius. A set of n locations non related with the user is rated. This rating is used to produce an ordered list of location recommendations.

The authors present two types of methods to rate locations. The first is to compute the cosine similarity between locations, and selects the first n most similar to the user's locations (i.e., locations annotating the user's photos). The second type refers to model based methods, which solve the problem of discovering M_1 and M_2 from a partially observed co-occurrence matrix, with $M_1 \in \mathbb{R}^{u \times k}$, $M_2 \in \mathbb{R}^{l \times k}$, l locations, u users, and k latent factors. Multiplying M_1 and M_2 gives an approximation of the complete co-occurrence matrix [46]. This study explores an important aspect of geographic information on social networks, which is the ability to infer the future interest of a user in a location. See section 7.2 for our contribution to the topic.

Liu et al. [35] also present an hybrid sub-space learning method for image annotation on location based social networks. The method represents the semantic information of a location in two spaces, a visual and a textual space. The textual space relates location with keyword according to co-annotation patterns, and the visual space represents a relationship between keywords and visual features (from the image annotation patterns in the network). The visual space should enable the discovery of relevant locations for non-annotated images. The evaluation uses a Flickr dataset with 3 309 698 images, and reports an $F1$ score at rank 1 of 0.1366.

Guo et al. [22] present a framework for travel recommendation, which unifies relationships between users and locations. The framework performs a random walk with restart over the unified model to find good recommendations. This process promotes a set of locations where related users have visited, and locations related with this initial set. The framework is evaluated with a Flickr dataset of undisclosed dimensions, reporting a precision at rank 5 between 0.5 and 0.6 (reported in a chart).

Magnuson et al. [37] present a system for location based social network event recommendation, particularly on mobile network clients. The systems builds user profiles through sentiment analysis of opinions posted on location and during events. These profiles are then used to support recommendations of future events.

Wen et al. [77] present a system for route recommendation, from location based social

network information.N From a set of keywords and geographic region, the system recommends routes based on measures of attractiveness (i.e., proximity to a popular point of interest), time constraints (i.e., preferred visiting hours at each point of the route), and its relationship with influential users.

Zhuang et al. [85] present a method for identifying if a city is familiar to a user. This goal is different from discovering known locations because a city has many regions, and familiarity has a subjective meaning. This work is mainly focused of identifying interesting locations in a city where user could take photos from. To discover those places, the method identifies users familiar with the city. The method relates information from three different models: a social network model, which determines that a city is familiar to the user if the relationships of the user are familiar with the location; a time driven model, which determines that if the user has frequently been in a city, then is familiar with that city; and a location driven model, which determines that the user is familiar with a city after visiting different relevant regions of the city. The method is evaluated with a dataset of 937 633 Flickr users, obtaining positive results (reported in receiver operating characteristic curves).

Apreleva and Cantarero [1] present a method for location estimation for Twitter users. The method creates clusters of users based on their relationships, and propagates the location inferred for some users to the remaining users in their cluster. This approach targets low density graphs, and not the entire network.

Bao et al. [3] survey the state of the art of location based social network analytics. The authors review work on location based recommendation and prediction. The conclusions show that the demand for location based recommendation is increasingly focused on the current location of the user, which adds a spatio-temporal dimension to the problem.

The authors identify the problem of predicting the future location of a social network user as a main research challenge. The increasing number of specialized networks change the collaborative, spatio-temporal context of the domain, and are currently the topic of several research efforts.

2.3.3 Discussion

Content, user, keyword, and location recommendations rely on the relationships between network elements, which generally enable the definition of relevance metrics between those elements [18, 33, 67].

Several authors focus present hybrid approaches to the automatic annotation problem, which extend the traditional content analysis approach with a collaborative filtering methods [27, 35, 42]. An affective approach to the problem is the use of machine learning methods, modeled with the information on the network [27, 35].

Presently, the active research topics on location based social network research are user location inference, whether to estimate [1, 44, 45, 85] or recommend locations to the user [22, 37, 76, 77], recommend location annotations for item [23, 27, 35, 46], or

improving on keyword recommendations based on their relationships with geographic annotations [73, 82].

Spatial information on social network is closely related with the spatio-temporal activity of the users [3], and is becoming related with their annotation efforts, which creates opportunities to answer questions such as whether an user, with a set of interests (which we identify from annotations), will be at (or be interest in) a place, for a specific time period.

2.4 Methods and Tools

To evaluate our proposals we rely on a set of tools for graph sampling and machine learning. We also present an interaction and visualization framework for productive networks. This sections presents the most relevant work related with our contributions.

2.4.1 Graph Sampling

Leskovec et al. [31] a set of characteristics present in social networks' graphs, which led to the definition of a graph generation model, the Forest Fire model. This model later inspired a network sampling algorithm [30] which produces a graph preserving the same growth properties of the original graph, which are:

Densification power law The number of edges increases over time at a rate described by a power law of the number of nodes, i.e., $e(t) \propto n(t)^{\alpha}$, where $\alpha > 1$, with $e(t)$ edges at time t, and $n(t)$ nodes at time t.

Shrinking diameter The graph diameter reduces over time.

The Forest Fire model, proved to be particularly suitable to sample social graphs. The approach follows a pattern in which highly linked nodes have an improved chance of being reached by new nodes.

2.4.2 Support Vector Machines

A Support Vector Machine – SVM – is a supervised learning model which assigns data items into one of two categories [9]. The model is built around a training set of examples of both categories.

The SVM is fitted to a sample of elements to classify. This fitting process is executed with a training set, with examples of correctly classified data. Ideally, this set is balanced with elements of both possible classifications.

The complete set of correctly classified data is the *ground truth*. Part of this ground truth should be used to train the classifier, and the remaining should be used to evaluate its performance.

Usually, the data contains several feature, which are the dimensions used to infer the classification model. The success of a classifier is determined by its training conditions, i.e., the set of features used to infer data patterns and the training set.

The challenge is twofold:

- Determining if the training set accurately represents the universe under study (see section 2.4.1);
- Selecting a robust set of element features.

2.4.2.1 Formal Definition

Support vector machines solve classification problems [4], which, in the context of this thesis, we assume to be a two-class classification problem, whose classic formalization is a linear discriminant function such as

$$y(x) = w^T \phi(x) + w_0$$

where $x \in \mathbb{R}^n$ is a feature space vector, $y(x) \in \{-1, 1\}$ a function of feature space vectors, $w^T \in \mathbb{R}^n$ parameters in feature space domain, which are determined iteratively by the learning process, $w_0 \in \mathbb{R}^n$ a bias parameter, and $\phi(x)$ a non-linear transformation which represents information in a higher dimension space.

Our model provides the framework to chose which features of a productive network may be used to build classifiers. The classification analysis phase provides insight about which features should be used to obtain good results.

Feature spaces are represented by $\mathcal{F}$, which is determined by the cardinalities of the feature sets $\mathcal{F}_1$ to $\mathcal{F}_n$, such that:

$$\mathcal{F} = \langle |\mathcal{F}_1|, ..., |\mathcal{F}_n| \rangle$$

Not all features in the data are relevant to the classification process. We follow a trial and error approach, guided by a general heuristic, to determine the best set of features to use in our experiments.

2.4.3 Network Visualization

Namata et al. [40] argue that zooming, filtering, clustering and layout techniques are useful to deal with large datasets, but are still limited in the number of attributes they can display at one time and do not allow comparisons of different subsets and aspects of the data. The authors propose a framework, constituted by a set of guidelines for network visualization and interaction tools, aiming at multiple interactive, coordinated views of the same network. According to this framework, tools should enable the selection of subnetworks for analysis, and provide specific visualization metaphors to represent different types of data (e.g., tree-maps for tree subnetworks, or node-link diagrams for

larger collections). One final guideline states that tools should enable links between different network views, in which the selection of an element in one view should enable the navigation to other views containing the element (or parts of it).

The authors implement the *DualNet* tool to evaluate their framework, and evaluated it using a subset of the Enron email collection (from 2000 to 2001) in a user trial, with positive results. Figure 2.1 presents sample views of the evaluation dataset.

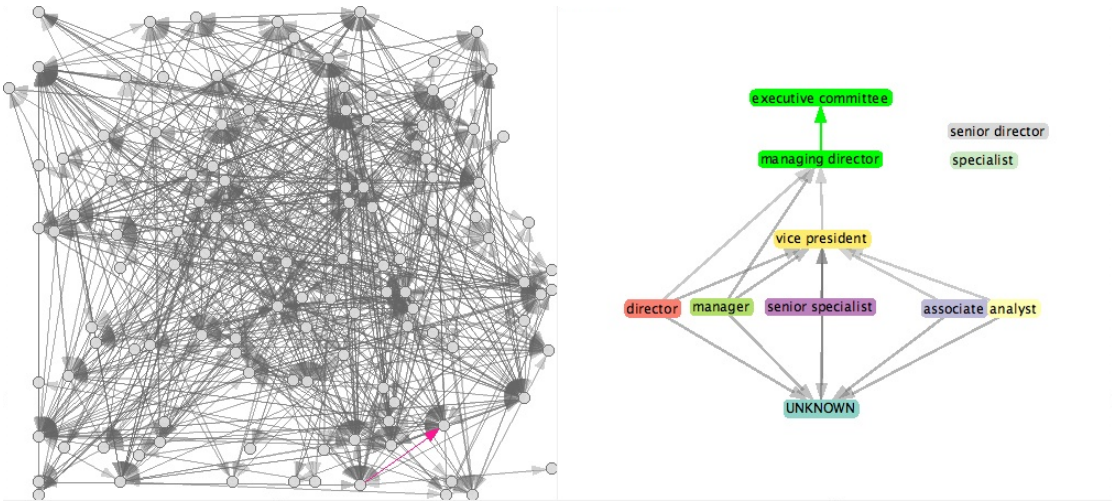

Figure 2.1: Title nodes shown in a tree to lines of authority (right) and email communication graph (left), separated using a force directed layout, and an highlighted upper management communications [40].

We rely on the notion of linked views in our visualization tools, to help users navigate on a dataset, enabling discovery, while keeping the context clear and the data quantities manageable. See chapter 7.1 for details on our approach to network visualization and interaction.

2.5 General Discussion

This chapter presents an overview of the research topics covering out research questions. These are the social and collaborative networks organization and characterization, annotation and geo-referencing systems, and content, user, keyword, and location recommendations.

Social network services are increasingly popular, and have a tendency to become specialized [19, 69]. Most network exhibit a similar graph structure [31, 41], and users tend to assume specific roles [5], according to the social interaction dynamics supported by the network [24, 43], and the communities they belong to [2, 65, 84].

Our work focus on the keyword and geographic annotation systems to discover relationships that are hard to identify. There are several use patterns for annotations, which vary from unregulated [81] to fixed taxonomy [8, 86]. In the former case, keywords need to be discovered, however phenomena like folksonomies emerge [39]. In the latter, the

users do not create keywords, but the annotation may suffer from a larger gap between the semantic the user intended and the one that is available [11]. Annotation systems and keyword categorization are key aspects of our research problem, given its role as link between network elements.

Our case studies include Flickr and Twitter (see section 6.2), which are very popular for method evaluation [1, 2, 11, 23, 24, 27, 28, 34, 43, 44, 45, 73, 81, 85]. Generally, good results with these datasets hint towards good results with other networks. The main reasons are:

- Users are able to annotate content without restrictions;
- Users have a low annotation activity.

We are influenced by recommender systems, which rely on methods that discover information. As previously stated, our methodology is designed to support the recommendation of users, keywords and locations.

The geographic dimension is appearing in many modern networks, which enables a new type of features, such as location recommendation or prediction.

The contributions presented in this chapter show that discovering keywords, locations, and content items on social network services is usually addressed when designing automatic annotation systems, or user, keyword, location and content recommendation systems.

Usually, the challenge is to find the best suited keyword to describe an item, the best suited location to a given query, or the best suited user for a professional task. Instead of answering a direct query, we are interested in discovering potential interests of the user. We know that these interests are represented in the annotation system, and our goal is to model the network in a way that enables this type of discovery.

Chapter 3

Productive Networks

Summary

This chapter presents a productive network study, which surveys a large collection of social network, content sharing, and content indexing services, enabling the formal definition of our domain. We propose to analyze the elements and operations that these networks offer to support the user's goals, and show that there are sufficient common aspects to support a generic productive network model.

The term *social network* is actually used to describe different types of networks, some of which are not explicitly focused on the definition of social network (see chapter 1). We propose the term *productive network*, which is the fundamental concept supporting the models and methods presented in this thesis. It refers to the common structural aspects found in social networking services.

To answer **RQ 1**, which queries about the common information concepts shared by different types of network, we designed a study whose goal is to systematically collect evidence of this common structure, which is a fundamental step towards a theoretical model of productive networks.

3.1 Definition and Subtypes of Productive Networks

As presented in chapter 1, we propose three terms to categorize network services: *Social Networks*; *Content Sharing Networks*; and *Content Indexing Network*. The individual definition of which are:

Social Network (SN) is a system which aims to replicate the network behavior through which human beings relate which each other. Examples are Facebook[1] and Google Plus[2];

Content Sharing Network (CSN) is a system where the main goal is to host and make available the content uploaded by users. We include in this category networks through which communities emerged, and in which some aspects of a social network behavior happen. Examples are Flickr[3] and Instagram[4];

Content Indexing Network - CIN is a system mainly focused on enabling the search for content that may be hosted somewhere else, or that is not freely available. The goal is to create awareness that such content exists. Examples are the ACM Portal[5] or the IEEE Explore[6].

To guide our study, we propose two dimensions to classify networks: *discovery focus* and *upload policy*. The discovery focus is either *users* or *content*, with most networks actually enabling both, while tailored to promote one over the other. The upload policy determines if the users are able to freely upload content or if this is regulated by a curation process. Figure 3.1 illustrates the classification of the three types of network according to these dimensions.

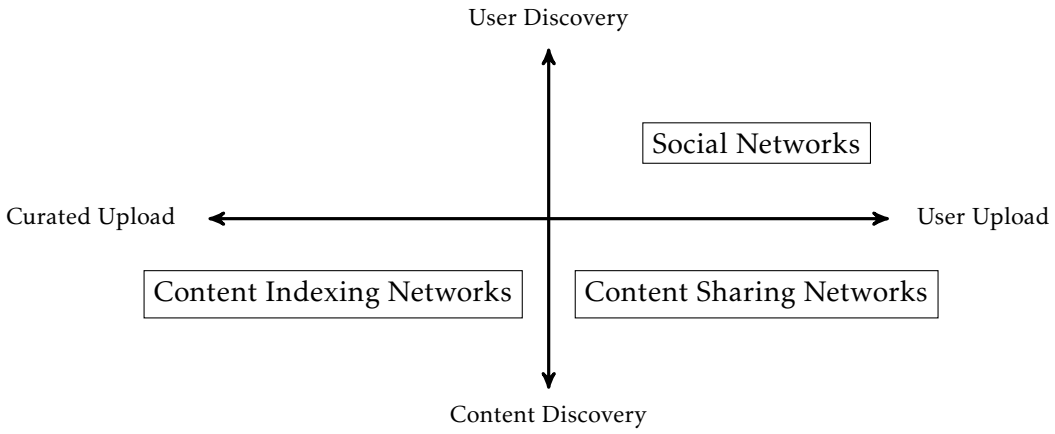

Figure 3.1: Types of productive networks according to upload policy (horizontal axis) and main discovery focus (vertical axis).

We propose to name the set defined by these three types of network as *productive networks*. We coined the term *productive network* to shift the focus of our definition from the social aspect of user relationships and organization, particularly focused by some networks, to the evidence of human production, found in all networks of our study.

[1] https://www.facebook.com
[2] https://plus.google.com
[3] https://www.flickr.com
[4] https://www.instagram.com
[5] http://portal.acm.org
[6] http://ieeexplore.ieee.org

3.2 Study Methodology

The aim of the study is to organize evidence to support a model for productive networks. We consider an initial set of networks (described in section 3.2.1), which is the baseline for the identification of similar networks, relevant for our study.

All networks considered candidates to be included in the study were preliminarily evaluated according to the following rules:

- Networks that enable searching for items or users;
- Networks with free registration.

Networks enable annotation through different user interfaces which differ considerably between services. These differences may be motivated by the content media type, network scope, or particular design guidelines, which may include hiding the annotation system on free textual descriptions of the content.

The study systematically describes and categorizes networks according to the following dimensions: supported user relationships; content media type; context and policy for annotation keywords; context and policy for spatial annotation of content; focus of the search tool. These dimensions are defined such as:

Supported user relationships Users may relate with individual users or with groups.

Content media type Media type may be *text*, *image*, *video*, or *url* (a particular case of text).

Context and policy for annotation keywords Annotation keywords may be used to describe or classify content. Keywords may be part of the content, or separate from it. The network may provide a specific taxonomy for keywords, and users may or not be able to create new keywords.

Context and policy for spatial annotation of content Content may be annotated with specific locations, which may be represented by specific coordinates or with the added semantic of geographical places. Locations may refer to points or areas, and be part of the content or separate from it.

Focus of the search tool The network may provide search tools for users, items, keywords, or locations.

3.2.1 Network Seed

The initial set of networks (the *seed*) was composed of the top 16 networks listed at Wikipedia's *Social Networking Websites List* [7], on january 1, 2013. The list was ordered by

[7] https://en.wikipedia.org/wiki/List_of_social_networking_websites

the Global Alexa Page Ranking [8], and by declared registered number of users. Table A.1, in appendix A, presents all networks included in the study, and both ranking values. Networks of the initial set are signaled as seed.

3.3 Studied Networks

To the initial set of 16 networks, we added 25 related networks, to a total of 41 productive networks, using to the following criteria (all networks in the examples are described in section 3.3.1):

- Reviewing a network in the set identifies another network as a key representative of its main functionality, e.g., Flickr suggested Picasa, and Facebook suggested VK;
- Reviewing a network in the set identifies a functionality which is addressed by another network, using a different media, e.g., Flickr suggested Youtube, LiveJournal suggested publication sharing networks, such as ACM Portal;
- Combining two domains of applications, e.g., we searched for social networks for scientific publications and identified ResearchGate.

Each network is categorized according to its subtype, which may be: Content Indexing Network (CIN), Content Sharing Networks (CSN); and Social Network (SN).

3.3.1 Content Indexing Networks

Academia.edu (www.academia.edu) CIN Academia.edu is a social network for academics, that relates users with their scientific publications. Users are also related with a set of keywords that describe their interests. Users register their publications, which are available for query. Query results are displayed with the user's keywords, enabling navigation through the keyword graph. Publications may also display information about the city or country of origin.

ACM Portal (portal.acm.org) CIN The ACM Digital Library is a research database indexing scientific publication, mainly from computer science journals and conferences. Each publication has a set of authors and keywords. Most ACM publications enforce a taxonomy of keywords, the ACM Computing Classification System. The system provides search by keyword and retrieves all publications annotated by the term. Publications may also display information about the city or country of origin.

Arxiv (arxiv.org) CIN Arxiv is a research database indexing preprints of scientific publications from electrical engineering and computer science conferences and journals. Each publication has a set of authors and keywords. Some publications enforce a taxonomy

[8] http://www.alexa.com/topsites

of keywords, some leave it entirely to the authors, and some enable both approaches by providing two sets of keywords per publication. The system provides search by keyword and retrieves all publications annotated by the term. Publications may also display information about the city or country of origin.

IEEE Explore (ieeexplore.ieee.org) CIN IEEE Xpplore is a research database indexing scientific publications from electrical engineering and computer science conferences and journals. Each publication has a set of authors, keywords, and may have some location description. Some publications enforce a taxonomy of keywords, some leave it entirely to the authors, and some enable both approaches by providing two sets of keywords per publication. The system provides search by keyword and retrieves all publications annotated by the term.

Mendeley (www.mendeley.com) CIN Mendeley is a scientific bibliography organizer, published by Elsevier. Users may register documents that they authored, or which they are interested in. Entries may contain location descriptions, referring to the author or publication event. The system uses the user content to enable a search for documents. Each document is registered with its set of authors, and user generated keywords.

Science Direct (www.sciencedirect.com) CIN Elsevier Science Direct is a research database indexing the scientific publications by Elsevier. Each article has a set of authors and keywords, which may include location descriptors, and the system provides the refinement of search results through keywords.

Slashdot (slashdot.org) CIN Slashdot is a news aggregation service which enables users to submit urls to news articles. Submissions are evaluated by curators, who decide whether it fits some of the services interests. Each entry is posted with a comment by the user, and is annotated according to a fixed taxonomy. Users are free to comment entries, and these comments are rated by curators. The service provides search by entry title and keyword.

Springer (www.springer.com) CIN The Springer library is a research database indexing scientific articles and books published by Springer. Entries have (article or book) have a set of authors and keywords, and some also contain location descriptors. The system provides free text search.

Web of Science (webofknowledge.com) CIN Thomson Reuters agency Web of Science service is a scientific publications index. It does not host the publications, but maintains a cross referencing graph between the scientific production. Each entry is annotated according to its publication keywords, and some contain location descriptors. The system provides free text search.

Yelp.com (www.yelp.com) CIN Yelp.com is a business indexing service, which enables customers to rate and review their experience. The system supports two types of user:

businesses and customers. Businesses may georeference and annotate their establishments according to a taxonomy. The system provides free text search.

3.3.2 Content Sharing Networks

Allrecipes (allrecipes.com) CSN Allrecipes is a recipe sharing network. The service enables registered users to rate and comment recipes, which are annotated according to a fixed taxonomy. With an optional paid registration, users may connect with each other and propose changes to other users' recipes. The service provides search for recipes by keywords and users. Some recipes may also display information about the city or country of origin.

Asana (www.asana.com) CSN Asana is a collaboration support system, focused on email replacement. The system enables users to create collaborative projects, contextualized by workspaces. Work items may be freely annotated. The system provides free text search.

Blogger (www.blogger.com) CSN Blogger is a blogging system in which users publish multimedia content, and may annotate content items with user created keywords. Users may subscribe for alerts about updates on other users' blogs. Some annotations refer to locations. The system provides free text search.

Del.ici.ous (delicious.com) CSN Del.ici.ous is a service which provides users to freely annotate and share bookmarks to websites. Users may follow other users. The services promotes the integration with other networks so that users may discover and follow their friends.

Flickr (www.flickr.com) CSN Flickr is an image and video hosting service. Users share media that is available through the website, and also accessible from other social media websites and blogs. The content is annotated with unrestricted keywords, without taxonomy, may be georeferenced, and can be public or private. Keywords are available for system-wide content search. It is also possible to use keywords to filter a search over the user content.

Github (github.com) CSN Github is a source code sharing service, supported by the git distributed version control system. Users create code repositories, which may be cloned by other users. The systems provides free text search, filtering results by programming language and activity metrics.

Goodreads (www.goodreads.com) CSN Goodreads is a content sharing network, which enables users to post updates on reading preferences. Users rate and review literary works. The system provides search by book author, title and ISBN. The annotation of literary works may include location references.

Instagram (instagram.com) CSN Instagram is an image and (short, 15 video seconds) video hosting service. Users share media through mobile phone clients. The content is

annotated with unrestricted keywords, without taxonomy, and is available to everyone, or only to a group of users. Some images are annotated with geographic locations. Users may also describe themselves with a set of keywords. Keywords are available for system-wide users and content search

Last.fm (www.last.fm) CSN Last.fm is a music recommendation service, which also supports a social network. The system streamed music until 2014, and is now focused on the recommendation service, while displaying musical performances posted on video sharing services. Users may enter groups with a common interest. Authors and songs are annotated with user created keywords. The system provides free text search.

LiveJournal (www.livejournal.com) CSN Livejournal is a blogging service. Users post multimedia content, which may be annotated with user created keywords. Some annotations contain location descriptors. User may follow other users, and grant access to private content. The system relies on Google custom search service [9] to support free text search.

Picasa (picasa.google.com) CSN Picasa is an image organization package. The user interface is a desktop application, integrated with a website (picasaweb.google.com). The system is also integrated with Google Plus. Users share media through the Google Plus integration and the Picasa website. The content is annotated with unrestricted keywords, which may contain location descriptors, and the system guides the users to annotate human faces with names. Keywords are available for system-wide content search. It is also possible to use keywords to filter a search over the user content.

Pinterest (www.pinterest.com) CSN Pinterest is media sharing service where users may post content and bookmark urls to interesting articles. The content is shared through boards. Unrestricted keywords, marked with a number sign, are available to annotate each post, which may refer to geographic locations, and users may also describe their content freely. The system provides search by keyword, and retrieves results which are annotated with the keyword or which include the keyword in the description.

Scridb (www.scribd.com) CSN Scridb is a books and documents sharing system. Users have access to edited books and user submitted work, both associated with a description. Books and documents are annotated with keywords governed by a taxonomy managed by the system. The systems provides a system wise search by keyword and text present in the item's description.

Slack (slack.com) CSN Slack is a collaboration support system, which enables users to upload multimedia documents. User interaction is centered around a messaging platform, which enables users to annotate content on different chat rooms. The system provides free text search.

[9] https://cse.google.com/cse

Trello (trello.com) CSN Trello is a collaborative task management systems. Users create task boards, each with several task lists. Tasks may be freely annotated with labels. The system provides free text search.

Tumblr (www.tumblr.com) CSN Tumblr is a blogging platform which enables users to follow and comment on blog posts. Users publish multimedia items, which can be annotated with keywords. Several posts contain location information. Users may also follow blogs of other users. The system provides free text and keyword based search.

Vimeo (www.vimeo.com) CSN Vimeo is a video sharing service. Users may freely annotate videos and follow other users. Videos may be annotated with location descriptors. The system provides search by title, user, and keywords.

Wordpress (www.wordpress.com) CSN Wordpress.com is a blogging system that uses the Wordpress platform. Each user publishes multimedia items on their blog, which may be annotated with keywords, and location descriptors. Users may subscribe for alerts about updates on other users' blogs.

YouTube (www.youtube.com) CSN Youtube is a video sharing service, which enables users to freely annotate their videos. Some users and videos include location information in their descriptions. Users may also rate and comment other users' videos. The system also enables a user to follow other users. The system provides search by title, user, and keywords.

IMDB (www.imdb.com) CIN The Internet Movie Database (IMDB) is a motion pictures, television and web shows indexing service, which enables users to rate and review entries. While the registration and annotation of each entry is curated, rating and reviews are freely contributed by registered users. Some annotations refer to locations related with the entry. The systems provides search by title, description, and character and cast names, while supporting browsing by genre.

3.3.3 Social Networks

Endomondo (www.endomondo.com) SN Endomondo is a social network which enables users to upload their performance on various sporting activities. Each item represents a georeferenced activity, which must be annotated according to a taxonomy. The network supports private groups of users.

Facebook (www.facebook.com) SN Facebook is a social network, which enables users to share media content and urls. Unrestricted keyword annotations are available since 2013, and are available search of content accessible to the user. Some items are georeferenced. The system supports free text search for content and users.

Foursquare (www.foursquare.com) SN Foursquare, and its check-in service Swarm, is a location-based social network which enables users to announce that they are in a specific

location, and register comments and ratings about locations of interest. The user queries the system for locations of business descriptions, and can choose where to go based on previous comments and ratings.

Google Plus (plus.google.com) SN Google Plus is a social network that relates users with each other through circles, which represent different contexts of relationships. Users may freely annotate content, and some items are georeferenced. The system provides free text search.

Hi5 (www.hi5.com) SN Hi5 is a social network, which enables users to share media content and urls. Users may freely annotate content. The system provides free text search, and some items are annotated with location references.

LinkedIn (www.linkedin.com) SN LinkedIn is a social network marketed for business and professionals. Single users register items that describe their education and training, past experience and current occupation. Organizations register their description and job offers. All content can be described by keywords. The system provides search for job offers and companies.

MySpace (www.myspace.com) SN MySpace is a social network, focused on music discovery and artist promotion. User post multimedia updates, follow and are followed by other users, and may freely annotate content. Some annotations contain location descriptors. The system provides free text search.

ResearchGate (www.researchgate.com) SN ResearchGate is a social network, which enables users to post updates on their scientific production. Each entry is annotated with the publication keywords, and may include location descriptors. Relationships and communities are contextualized by common research interests or scientific publications. The system provides free text search.

Runkeeper (runkeeper.com) SN Runkeeper is a social network focused on sporting activities and health data. Items represent georeferenced activities or health measurements. Each sporting activity must be annotated according to a taxonomy. The network supports private groups of users, which generates an enhanced social graph with health information.

Twitter (www.twitter.com) SN Twitter is a micro blogging platform which support short posts with text or images. Users share content by following the content update feed of each other. Posts can include unrestricted keywords, marked with a number sign (#), which support a system-wise search of posts. Twitter's number sign notation influenced both Facebook and Google Plus approaches. Some posts are georeferenced.

VK (vk.com) SN VK (originally *VKontakte*) is a social network, which enable users to post social updates and share multimedia content. Posts may be georeferenced. Users are

able to establish relationships and create groups. The content may be freely annotated. The system provides free text search.

3.4 Results

This section presents a summary description of the data, and identifies the relevant set of evidence which motivates the productive network model presented in chapter 4. The complete study dataset is presented in appendix A.

Figure 3.2-A presents the content media type for all types of network, showing a wide range of media types. Given our focus on meta-data analysis, these results suggest a variety of contexts where our indirect relationship identification methods may be implemented.

The keyword policy and use, presented in figure 3.2-B, show that keywords are used for the purposes of content description and classification simultaneously. Figure 3.2-C presents the location policy and use, showing that most networks enable the annotation of content with place references.

The evidence produced by the survey enables the definition of a set of statements that summarize our findings. The first statement is that every *productive network has a representation of user, content item, and annotation keyword*.

There are two contexts where a person's identity, or alias, may appear in a network: as an author of content; or as the user who shared that content. For the purposes of representing users and personal interests, we consider that *users have an ownership relationship with content items, which they may have or not authored*.

Table 3.1 presents results on keyword policy for the three types of network. While fixed taxonomies are associated with curated upload policies and content discovery networks, the ability to freely create keywords is available in the majority of the networks. All networks, either focused on content or user discover, enable search by keyword.

Table 3.1: Summary of the survey results describing the data set counts, and the differences between the seed and final set of networks. Results about keyword policy and use refer to the final set of networks. All results are clustered by network type.

	Count		Keywords		
	Seed	Survey	Free to create	Searchable	With taxonomy
SN	7 (44%)	11 (27%)	8 (72%)	11 (100%)	4 (36%)
CSN	8 (50%)	20 (49%)	18 (90%)	20 (100%)	9 (45%)
CIN	1 (6%)	10 (24%)	8 (80%)	10 (100%)	10 (100%)
TOTAL	16	41	34 (85%)	41 (100%)	23 (56%)

From the results in table 3.1, we conclude that *keywords are available for annotation, and establish relationships between content items*. This relationship implies that *users become associated with the keywords they use to annotate content*.

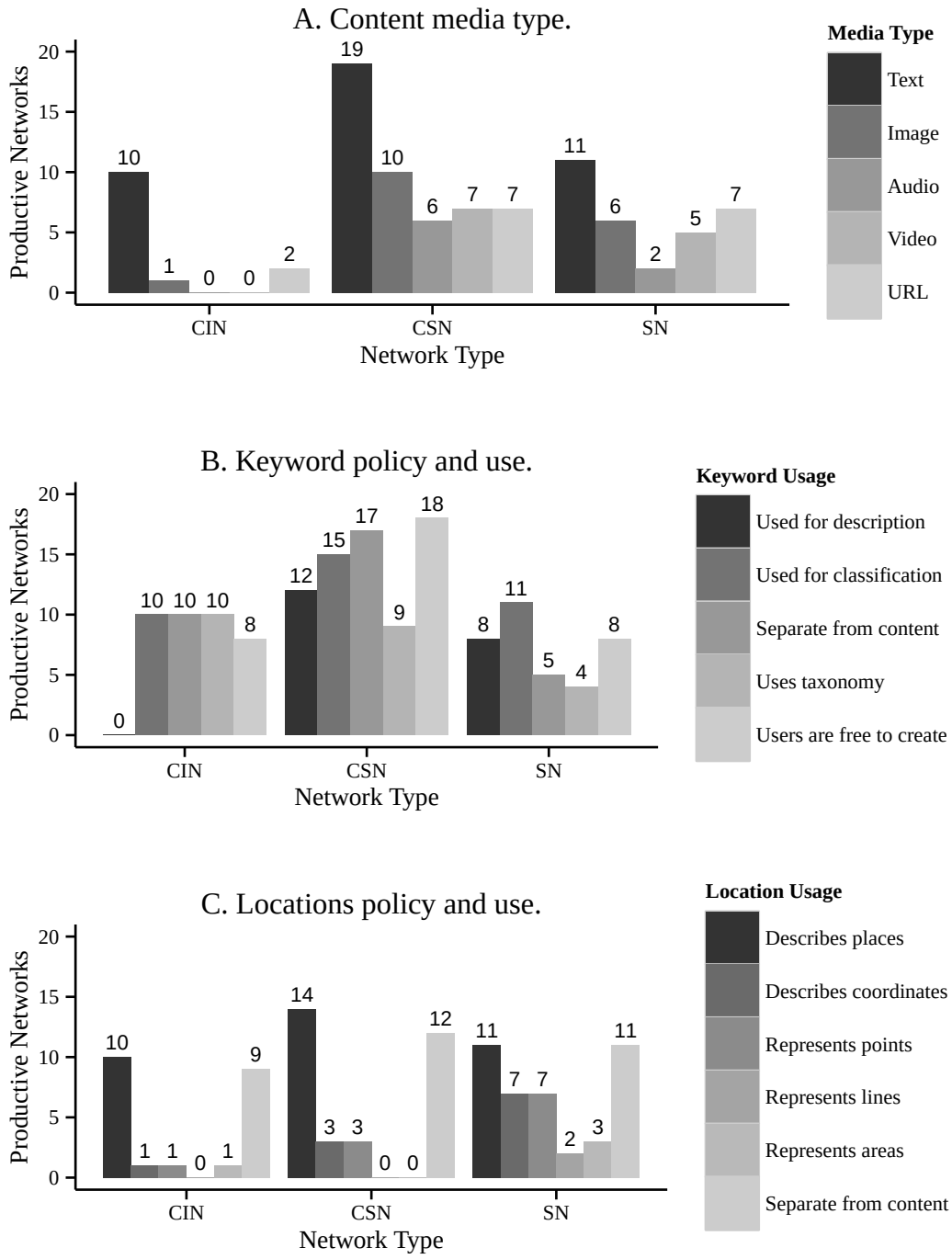

Figure 3.2: Productive network survey result summary, clustered by network type. There are 41 networks in the survey, with 10 CIN, 20 CSN, and 11 SN. (A.) Each network may support several media types. (B.) Keywords may be used to *describe* or *classify* content. They may be *separate from the content*, be regulated by a *taxonomy*, and/or the users may be *free to create* them. (C.) Locations may refer to a *place* or a only set or *coordinates*. Each location may describe a *point*, *polyline*, or *area*. Locations may also be *separate from the content*.

All networks enable search by keywords, which produces listings of content items and/or users. The relationships between users and content items, and between content items and keywords imply that *users that annotate content with the same keywords are related through those keywords*.

A subset of the networks in the survey enable support for the annotation of content items with locations. These annotations may refer to a specific point, described by coordinates, or to a place, which semantically may refer to a street, city or general area. We consider both the user provided annotations to describe locations, and locations inferred from a GPS sensor on the user device, which is available on some networks. Table 3.2 presents results related with location policy and use.

Table 3.2: Summary of the survey results about location keyword policy and use, presenting the differences between seed and final set of networks on location support. All results are clustered by network type.

	Supports locations		Describes places	Describes coordinates
	Seed (n=16)	Survey (n=41)		
SN	7/7 (100%)	11/11 (100%)	11 (100%)	7 (64%)
CSN	7/8 (88%)	14/20 (70%)	14 (70%)	3 (15%)
CIN	1/1 (100%)	10/10 (100%)	10 (100%)	1 (10%)
TOTAL	15/16 (94%)	35/41 (85%)	35/35 (100%)	11/35 (31%)

Table 3.2 implies that *some productive networks have a representation for locations which are used to annotate content items*.

All networks that support locations enable the annotation of content items with places, which provides semantic to the annotation. Locations in fact represent a specific case of annotation keyword, and enable the same type of relationships as keywords, therefore *users that annotate content with the same locations are related through those locations*.

3.5 Discussion

We are interested in systems which organize user production, and in evidence of a common set of information elements able to support a model for indirect relationship discovery. There is an increasing offer of social network services which, albeit being focused on different topics, do seem to share a common set of elements and operations.

We propose the term *productive network* to represent network services with these common aspects. Productive networks include three types of services, i.e., *social networks*, *content sharing networks*, and *content indexing networks*.

To systematically collect and categorize evidence of the common aspects of productive networks we conducted a survey. The survey results enable the identification of several statements which guide the definition of a productive network model, presented in chapter 4. Table 3.3 summarizes the evidence statements resulting from the survey.

Table 3.3: Evidence statements inferred from the productive network survey. ⋆ – statements regarding locations, which were inferred from a partial set of networks.

Evidence Statement	Description
E1	Productive networks have a representation of user, content item, and annotation keyword.
E2	Users have an ownership relationship with content items, which they may have or not authored.
E3	Keywords are available for annotation, and establish relationships between content items.
E4	Users are associated with the keywords they use to annotate content.
E5	Users that annotate content with the same keywords are related through those keywords.
E6⋆	Some productive networks have a representation for locations, which are used to annotate content items.
E7⋆	Users that annotate content with the same locations are related through those locations.

Statements in table 3.3 represent a list of requirements for both model construction and validation. Section 4.4 presents this validation.

Chapter 4

Productive Network Model

Summary

This chapter presents the definition of the productive network model. The concepts involved in the model reflect conclusions drawn from the survey evidence, presented in chapter 3. The model enables the representation of any productive network by a common set of elements, which will enable the application of indirect relationship discovery methods to each particular network and, eventually, to a group of networks.

We propose a productive network model, which has a representation of all statements of the evidence produced by the network survey, summarized in table 3.3 (see section 3.5).

Ultimately, we are interested in evaluating the model as a framework for indirect relationship discovery. This chapter presents the fundamental operations that support the indirect relationship discovery methodology (presented in chapter 5). The evaluation of both model and methodology is presented in chapter 6.

Following our research plan, the model was initially developed without artifacts for locations, which were later included as a specialization. Although the model is presented will all its elements, our research approach produces two distinct evaluations (see chapter 6).

This chapter addresses the second research question, **RQ 2**, which requires a characterization of the basic concepts and relationships of productive networks. It also addresses the fifth research question, **RQ 5**, which requires the characterization of locations and their relationships with the other elements of productive network.

The model refers to the concepts of *user*, *item*, *keyword*, and *location*, which we define such as:

User: is a user account on a network, generally uniquely identified by a user name or email address, representing a person or other entity; it has the ability to own items, and annotate them according to his preference, while subject to the network's annotation system's rules.

Item: is a content artifact, supported by a particular media, e.g., text, image, sound, or video.

Keyword: is a word, or set of words, which annotates items, subject to the network's annotation systems rules; it represents meta-information about items.

Location: is a special case of an annotation, representing a spatial location, corresponding to a spatial area on the world map.

The expressions in this chapter are presented using the notation described in Appendix C.

4.1 Formal Definition

The basic elements of the model are users, U, items, I, keywords, K, and locations, L. Items are owned by users, and annotated with keywords and/or locations.

Let us define U, I, K, and L such as:

$$U = \{U_1, \cdots, U_n\} \textit{ is a finite set of users, } n \geq 1$$
$$I = \{I_1, \cdots, I_m\} \textit{ is a finite set of items, } m \geq 1$$
$$K = \{K_1, \cdots, K_o\} \textit{ is a finite set of keywords, } o \geq 1$$
$$L = \{L_1, \cdots, L_u\} \textit{ is a finite set of locations, } u \geq 1$$

Note: The subscripts used in the definitions serve to distinguish between elements of the same set. We use i, j, k for users, p, q, r for keywords, t, u, v for items, l, g, h for locations, and m, n, o, u for set dimensions.

Definitions 1, 2, and 3, represent the basic item management operations that the network provides to its users.

Definition 1 *The ownership an item, $I_t \in I$, by one user, $U_i \in U$, is defined by:*

$$O(U_i) = \{I_u \mid I_u \in I \wedge I_u \textit{ is owned by } U_i\}$$
$$Own(U_i, I_t) \Rightarrow I_t \in O(U_i)$$

Definition 2 *The annotation of an item, $I_t \in I$, by a keyword, $K_p \in K$, is defined by:*

$$A(I_t) = \{K_q \mid K_q \in K \wedge K_q \textit{ annotates } I_t\}$$
$$Annotate(K_p, I_t) \Rightarrow K_p \in A(I_t)$$

Definition 3 *The geographic referencing of an item, $I_t \in I$, by a location, $L_l \in L$, is defined by:*

$$G(I_t) = \{L_l \mid L_l \in L \wedge L_l \text{ is associated with } I_t\}$$
$$GeoRef(L_l, I_t) \Rightarrow L_l \in G(I_t)$$

We refer to keywords that are used in annotations as the user's *direct keywords*. In definition 2, the set $A(I_t)$ is the set of direct keywords of item I_t.

Definition 4 *The set of all direct keywords, UK, of an user, $U_i \in U$, is defined by:*

$$UK(U_i) = \{K_p \mid K_p \in K \wedge \exists I_t \in O(U_i) : (Annotate(K_p, I_t))\}$$

We now define the relationships that items and keywords enable between users. We begin with the definition of direct relationship, which establishes a link between users.

Definition 5 *A direct relationship, DR, between two users, $U_i, U_j \in U$, is defined by:*

$$\underset{i \neq j}{}$$

$$DR(U_i, U_j) = \{K_p \mid K_p \in K \wedge \exists \underset{t \neq u}{I_t, I_u} \in I : (\begin{Bmatrix} Own(U_i, I_t), Own(U_j, I_u) \\ Annotate(K_p, I_t), Annotate(K_p, I_u) \end{Bmatrix})\}$$

Based on definition 5, we define keyword supported indirect relationships such as:

Definition 6 *A indirect relationship, IR, between two users, U_i and U_j, is defined by:*

if

$$DR(U_i, U_j) = \{\emptyset\},$$
$$\exists \underset{k \neq i \neq j}{U_k} \in U : \begin{Bmatrix} DR(U_k, U_i) \neq \{\emptyset\} \\ DR(U_k, U_j) \neq \{\emptyset\} \end{Bmatrix}$$

then

$$IR(U_i, U_j) = \{K_p \mid K_p \in K \wedge K_p \in DR(U_k, U_j)\}$$
$$IR(U_j, U_i) = \{K_p \mid K_q \in K \wedge K_q \in DR(U_k, U_i)\}$$

As a special case of annotation, locations also enable direct and indirect relationships. We begin by redefining the direct relationship, presented in definition 5, now based on locations instead of keywords. Definition 7 presents the set of location supported direct relationships.

Definition 7 *For a user, $U_i \in U$, the set of all direct locations, UL, of all of the user's items is defined by:*

$$UL(U_i) = \{L_l \mid L_l \in L \wedge \exists I_t \in O(U_i) : (GeoRef(L_l, I_t))\}$$

We are now able to define direct relationships based on locations, DRL, between two users, U_i and U_j, such as:

$$DRL(U_i, U_j) = \{L_l \mid L_l \in L \wedge \exists \underset{t \neq u}{I_t, I_u} \in I : \begin{Bmatrix} Own(U_i, I_t), Own(U_j, I_u), \\ GeoRef(L_l, I_t), GeoRef(L_l, I_u) \end{Bmatrix}\}$$

Finally, we define location supported indirect relationships, from location supported direct relationships, such as:

Definition 8 *An indirect relationship, IRL, between two users, U_i and U_j, based on locations, is defined by:*

if

$$DRL(U_i, U_j) = \{\emptyset\},$$
$$\exists U_k \in U : \left\{ \begin{array}{l} DRL(U_k, U_i) \neq \{\emptyset\} \\ DRL(U_k, U_j) \neq \{\emptyset\} \end{array} \right\}$$

then

$$IR(U_i, U_j) = \{\forall L_l \in L \mid L_l \in DRL(U_k, U_j)\}$$
$$IR(U_j, U_i) = \{\forall L_m \in L \mid L_m \in DRL(U_k, U_i)\}$$

4.1.1 Trivial Operations

The operation definitions presented in this chapter focus on the relationships between users and the other network concepts, with the objective of enabling the discovery of indirect relationships. There are several operations involved in the development of methods for indirect relationship discovery which we consider trivial, such as:

Obtain all items of a keyword: For a keyword, K_p, produce a list of all content items annotated with K_p.

Obtain all users of a keyword: For a keyword, K_p, produce a list of all users who annotate content items with K_p.

Obtain all items of a location: For a location, L_l, produce a list of all content items annotated with L_l.

Obtain all users of a location: For a location, L_l, produce a list of all users who annotate content items with L_l.

Obtain an rank ordered list of elements: Sort a list with rank values by total order, or reverse total order.

4.2 Graphs

The model definition enables de description of graphs implicitly defined by the network. Table 4.1 presents all graphs enabled by relationships between users, items, and keywords.

Table 4.1: All graphs that may be defined using the concepts of the model, each with an unique combination of node ($\mathcal{V}$) and edge ($\mathcal{E}$) sets.

$\mathcal{G}_{id} = \langle \mathcal{V}, \mathcal{E} \rangle$
$\mathcal{G}_1$: Users connected through their items.
$\mathcal{V} = \{U_i \mid U_i \in U \wedge \exists_{j \neq i} U_j \in U : (DR(U_i, U_j) \neq \{\emptyset\})\}$
$\mathcal{E} = \{I_t \mid I_t \in I \wedge \exists_{i \neq k} U_i, U_k \in \mathcal{V} : (Own(U_i, I_t) \wedge Own(U_k, I_t))\}$
$\mathcal{G}_2$: Users connected through keywords, which annotate their items.
$\mathcal{V} = \{U_i \mid U_i \in U \wedge \exists_{j \neq i} U_j \in U : (DR(U_i, U_j) \neq \{\emptyset\})\}$
$\mathcal{E} = \{K_p \mid K_p \in K \wedge \left\{ \begin{array}{l} \exists U_i \in \mathcal{V} : Own(U_i, I_t) \wedge Annotate(K_p, I_t) \\ \exists_{j \neq k} U_j, U_k \in U : K_p \in DR(U_j, U_k) \end{array} \right\}\}$
$\mathcal{G}_3$: Items connected through their common users.
$\mathcal{V} = \{I_t \mid I_t \in I \wedge \left\{ \begin{array}{l} \exists_{i \neq j} U_i, U_j \in U : Own(U_i, I_t) \wedge Own(U_j, I_t) \\ \exists K_p \in K : Annotate(K_p, I_t) \end{array} \right\}\}$
$\mathcal{E} = \{U_i \mid U_i \in U \wedge \exists_{t \neq u} I_t, I_u \in \mathcal{V} : (Own(U_i, I_t) \wedge Own(U_i, I_u))\}$
$\mathcal{G}_4$: Items connected through their common keywords.
$\mathcal{V} = \{I_t \mid I_t \in I \wedge \exists K_p \in K, \exists_{u \neq t} I_u \in I : (Annotate(K_p, I_t) \wedge Annotate(K_p, I_u))\}$
$\mathcal{E} = \{K_p \mid K_p \in K \wedge \exists_{t \neq u} I_t, I_u \in \mathcal{V} : (Annotate(K_p, I_t) \wedge Annotate(K_p, I_u))\}$
$\mathcal{G}_5$: Keywords connected through users which use them to annotate items,
$\mathcal{V} = K_p \mid K_p \in K \wedge \exists I_t \in I : (Annotate(K_p, I_t))\}$
$\mathcal{E} = \{U_i \mid U_i \in U \wedge \exists_{p \neq q} K_p, K_q \in \mathcal{V} : (K_p \in UK(U_i) \wedge K_q \in UK(U_i))\}$
$\mathcal{G}_6$: Keywords connected through common items.
$\mathcal{V} = \{K_p \mid K_p \in K \wedge \exists I_t \in I : (Annotate(K_p, I_t))\}$
$\mathcal{E} = \{I_t \mid I_t \in I \wedge \exists_{p \neq q} K_p, K_q \in \mathcal{V} : (Annotate(K_p, I_t) \wedge Annotate(K_q, I_t))\}$

The graphs in table 4.1 describe all possible contexts of relationship paths between elements of the network. For the context of indirect relationships between users, we are interests in the graph that connects users through keywords – $\mathcal{G}_2$. This graph excludes isolated users, which cannot be related with any other user through any keyword. The model does not consider elements in isolation.

The graphs implicitly defined by the network, presented in table 4.1, may also include locations as nodes and edges. Table 4.2 presents all possible graphs enabled by locations.

Table 4.2: Extension of the graphs presented in table 4.1, using the location concept. Each graph represents an unique combination of node and edge.

$\mathcal{G}_{id} = \langle \mathcal{V}, \mathcal{E} \rangle$
$\mathcal{G}_7$: Locations connected through users. $\mathcal{V} = \{L_l \mid L_l \in L \wedge \exists I_t \in I : (GeoRef(L_l, I_t))\}$ $\mathcal{E} = \{U_i \mid U_i \in U \wedge \exists L_l \in \mathcal{V}, \exists I_t \in I : (Own(U_i, I_t) \wedge GeoRef(L_l, I_t))\}$
$\mathcal{G}_8$: Locations connected through items. $\mathcal{V} = \{L_l \mid L_l \in L \wedge \exists I_t \in I : (GeoRef(L_l, I_t))\}$ $\mathcal{E} = \{I_t \mid I_t \in I \wedge \exists L_l, L_m \in \mathcal{V} : (GeoRef(L_l, I_t) \wedge GeoRef(L_m, I_t))\}$
$\mathcal{G}_9$: Locations connected through keywords. $\mathcal{V} = \{L_l \mid L_l \in L \wedge \exists I_t \in I : (GeoRef(L_l, I_t))\}$ $\mathcal{E} = \{K_p \mid K_p \in K \wedge \exists I_t, I_u \in I : (\begin{cases} Annotate(K_p, I_t) \wedge Annotate(K_p, I_u) \\ \exists L_l, L_m \in \mathcal{V} : \begin{cases} GeoRef(L_l, I_t) \\ GeoRef(L_m, I_u) \end{cases} \end{cases})\}$
$\mathcal{G}_{10}$: Users connected through locations. $\mathcal{V} = \{U_i \mid U_i \in U \wedge \exists U_j \in U : (U_i \neq U_j \wedge DRL(U_i, U_j) \neq \{\emptyset\})\}$ $\mathcal{E} = \{L_l \mid L_l \in L \wedge \exists U_i \in \mathcal{V} : \begin{cases} \exists I_t \in I : Own(U_i, I_t) \wedge GeoRef(L_l, I_t) \\ \exists U_j \in U : L_l \in DRL(U_i, U_j) \end{cases}\}$
$\mathcal{G}_{11}$: Items connected through locations. $\mathcal{V} = \{I_t \mid I_t \in I \wedge \begin{cases} \exists U_i, U_j \in U : U_i \neq U_j \wedge Own(I_i, I_t), Own(U_j, I_t) \\ \exists K_p \in K : Annotate(K_p, I_t) \end{cases}\}$ $\mathcal{E} = \{L_l \mid L_l \in L \wedge \begin{cases} \exists I_u \in \mathcal{V} : GeoRef(L_l, I_u) \\ \exists U_j \in U : L_l \in DRL(U_i, U_j) \end{cases}\}$
$\mathcal{G}_{12}$: Keywords connected through locations. $\mathcal{V} = \{K_p \mid K_p \in K \wedge \exists I_t \in I : (Annotate(K_p, I_t))\}$ $\mathcal{E} = \{L_l \mid L_l \in L \wedge \exists I_t \in I, \exists K_q \in \mathcal{V} : (GeoRef(L_l, I_t) \wedge Annotate(K_q, I_t))\}$

Graphs provide an interesting framework to explore the content of productive networks. Section 7.1 presents an interaction and visualization tool that represents productive network information using these graphs, which uses the operators of the model to enable the user to switch from a visualization context to another. User evaluation shows that it is a valuable tool to discover and understand the context of relationships.

Location supported graphs provide context for future work on population density variation detection, discussed in section 8.2. Section 7.2 presents preliminary results on that topic, including developments for the emergency management domain.

4.3 Indirect Relationship Discovery

In the context of the graphs in tables 4.1 and 4.2, indirect relationships refer to shortest paths of size two. While indirect relationships may be defined using an arbitrary distance in the graph, we use size two because it takes advantage of the behaviors of users in productive networks, such as community membership and tendency to specialize in the network. Given this aspects, we believe that there is a higher potential for successful indirect relationship discovery in close proximity to the user in the network graph.

Considering graph $\mathcal{G}_2$, which relates users through keywords, these paths are defined by one keyword used by the user – a *direct keyword* –, and one that is not – an *indirect keyword*. Corollary 1 describes the set of indirect keywords, which are used to define indirect relationships.

Corollary 1 *Definition 6 enables the definition of a set of indirect keywords, $\mathbb{IK}$, of a user, U_i, such that:*

$$\mathbb{IK}(U_i) = \{K_p \in K \mid \forall U_j \in U, U_j \neq U_i, \mathbb{IR}(U_i, U_j) \neq \{\emptyset\} : K_p \in \mathbb{IR}(U_i, U_j)\}$$

We may formulate a similar description using locations. Corollary 2 describes the set of indirect locations, which may also be used to define indirect relationships.

Corollary 2 *Definition 8 enables the definition of a set of indirect locations, $\mathbb{IL}$, of a user, U_i, such that:*

$$\mathbb{IL}(U_i) = \{L_l \in L \mid \forall U_j \in U, U_j \neq U_i, \mathbb{IRL}(U_i, U_j) \neq \{\emptyset\} : L_l \in \mathbb{IRL}(U_i, U_j)\}$$

With a set of indirect keywords or locations for a user, U_i, building an ordered list of users with whom U_i has indirect relationships is a trivial operation (see section 4.1.1).

Both corollaries represent the same principle for indirect relationships discovery: *to build a list of indirect relationships for a user, U_i, we should discover indirect keywords or locations for user U_i*. This is the theoretical framework for the indirect relationship discovery framework, presented in chapter 5.

4.3.1 Indirect Keywords Discovery

We propose to build the set of indirect keywords of user U_i, which enables the construction of a list of users that use those keywords but do not have a direct relationship with U_i, i.e., the indirect relationships.

We calculate the number of items associated with each keyword in the list of indirect keywords of user U_i. This value will be used to sort the list.

Definition 9 *For a user, U_i, the list of indirect keywords with rank values, $\mathbb{IK}_r$, is defined by:*

$$R(K_p) = \mid \{I_t \mid Annotate(K_p, I_t)\} \mid$$
$$\mathbb{IK}_r = \{\langle K_p, |R(K_p)| \rangle \mid K_p \in \mathbb{IK}(U_i)\}$$

The sorted list of indirect keywords is a reversed total order of $\mathbb{IK}_r$. The rank value for each keyword, as described in definition 9, is one approach among many. For instance, the total number of keywords that share an item with the indirect keyword could be used to calculate a rank value. We found that the number of items annotated with the keyword was best suited. Section 6.2 presents results of an experiment which compares these methods.

Ultimately, after the list of indirect keywords is found, the goal would be to deliver a list of indirect users. We focus on the validation of the ranked indirect keywords list, and discuss the results of a set of experiments that are designed to evaluate the suitability of this list. Furthermore, in section 6.1 we also designed an experiment which uses learning methods to discover indirect keywords. Chapter 6 presents the results of that experiment.

4.3.2 Indirect Locations Discovery

The approach to indirect location discovery is similar to the indirect keyword discovery.

Definition 10 *For a user, U_i, the list of indirect locations with rank values, $\mathbb{IL}_r$, is defined by:*

$$R(L_l) = \mid \{I_t \mid GeoRef(L_l, I_t)\} \mid$$
$$\mathbb{IL}_r = \{\langle L_l, |R(L_l)|\rangle \mid L_l \in \mathbb{IL}(U_i)\}$$

Section 6.3 presents results of experiments which attempt to discover indirect locations.

4.4 Validation

The model concepts and operators should address the complete set of requirements summarized in table 3.3, in section 3.5.

Table 4.3: Cross reference between evidence statements inferred from the productive network survey and the model definitions.

Evidence Statement	Model concept and operators
E1	Definitions of the sets U, I, and K.
E2	Definition 1: the operator Own.
E3	Definition 2: the operator $Annotate$.
E4	Definition 4: the user keywords set, UK.
E5	Definition 5: the direct relationships set, DR.
E6	Definition of set L. Definition 3: the operator $GeoRef$.
E7	Definition 7: the direct relationships set, DRL.

Table 4.3 presents a cross reference between model definition and evidence statements. The definitions of indirect relationships are not included in the table because they are

a consequence of evidence based operators, and are not explicitly based on evidence statements. These definitions will be addressed in the model evaluation, presented in chapter 6.

4.5 Discussion

The productive network model provides the tools to describe indirect relationships. It is supported by the evidence collected by the survey in chapter 3.

The definitions of indirect relationships show that these are supported by indirect elements, i.e., keywords and locations. In fact, we have showed that it is trivial to discover indirect relationships from indirect elements. The challenge is in the identification of indirect keywords and locations.

The model enables the description of indirect keywords and locations which support the methodology presented in chapter 5.

The construction of the model addresses **RQ 2.**, which asked for a characterization of the relationships between productive networks information concepts. It also provides descriptions and fundamental operators required by all subsequent research questions.

Chapter 5

Indirect Relationship Discovery Methodology

Summary

This chapter presents the Indirect Relationship Discovery Methodology, which is supported by the productive network model, described in chapter 4. We focus on machine learning methods, which rely on systematic feature definitions enabled by our model. These methods were designed to facilitate the systematical evaluation presented in chapter 6.

The methodology addresses research questions **RQ 3**, **RQ 6**, and **RQ 4**, which ask for methods to identify indirect artifacts.

5.1 Methodology Outline

The productive network model enables several representations of the network information, providing different perspectives on the data and its relationships. In this chapter, the methodology is presented using the perspective/representation where users are connected through keywords or locations.

Considering graph $\mathcal{G}_2$ (users related through keywords, as described in table 4.1), a relationship refers to a path between two users. Similarly, considering graph $\mathcal{G}_{10}$ (users related through locations, as described in table 4.2), a relationship also refers to a path between two users.

Therefore, considering the definitions of direct relationship (see definitions 5 and 7) and indirect relationship (see section 4.3), these are represented in the graph such as:

- Direct relationship: is a path in the graph that connects two users that share a common keyword or location;
- Indirect relationship: is a path in the graph that connects two users through an existing path with two different keywords or locations, used individually by the users.

Figure 5.1 illustrates an indirect relationship (full line), and the direct relationship (dashed lines) that support it.

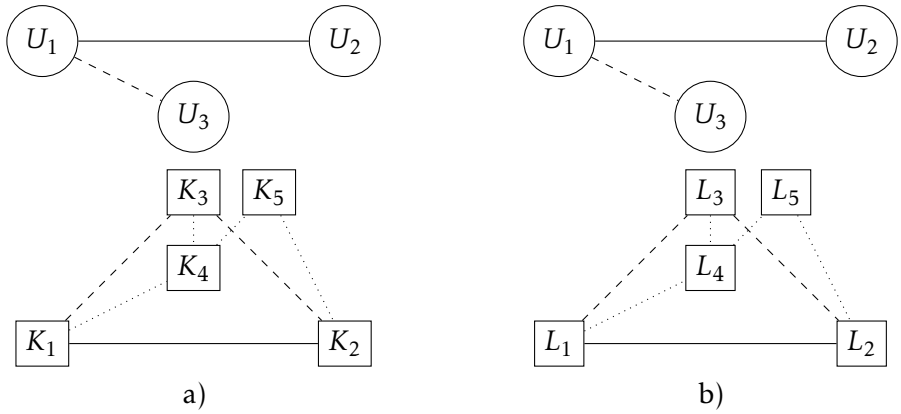

Figure 5.1: (a.) User U_1 uses keyword K_1, user U_2 uses the keyword K_2, and user U_3 the keyword K_3: the indirect relationship (full line) between users U_1 and U_2 is supported by the paths (dashed lines) between keywords K_1 and K_3, and keywords K_3 and K_2. Dotted lines represent other paths in the graph. (b.) Similar relationship, supported by locations.

Our methodology is focused on the discovery of the indirect keywords and locations that support indirect relationships. We refer to these as *indirect artifacts*.

Section 4.3 states that it is trivial to obtain indirect relationships from indirect artifacts, therefore the methodology goal is to build suitable ranked lists of indirect elements.

Our methodology development approach has two phases: a frequency analysis, which builds a ranked list indirect elements, observing all artifacts which verify the conditions of indirect element; and a classification analysis, which relies on machine learning methods to identify indirect artifacts.

5.1.1 Frequency Analysis

The frequency analysis phase is, actually, a first approach to solving the problem of identifying indirect artifacts. It provides insight about the relationships in the network, which helped guide the development of a more sophisticated approach, the classification analysis.

We present the frequency analysis in the keyword discovery context. Algorithm 1 presents the procedure to find the indirect artifacts for every user, from an existing dataset.

Algorithm 1 Frequency analysis. Obtains all indirect keywords of the user, from the relationships between her items and keywords. The *appendUnique* function appends the parameter list to the caller, without repetitions.

Require: $O(U_i) \neq \{\emptyset\}$
Require: $T'(U_i)$: a list of keywords of the user
1: $\mathbb{IK} = \{\emptyset\}$
2: **for all** $I_t \in O(U_i)$ **do**
3: **if** $T'(I_t) \neq \{\emptyset\}$ **then**
4: $I_{K_p} = list()$
5: $\mathbb{IK}_{K_p} = list()$
6: **for all** $K_p \in T'(I_t)$ **do**
7: **for all** $I_r \in I$ *such that* $I_r \notin O(U_i)$ **do**
8: **if** $K_p \in T'(I_r)$ **then**
9: $I_{K_p}.append(I_r)$
10: **end if**
11: **end for**
12: **for all** $I_u \in I_{K_p}$ **do**
13: **for all** $I_v \in O(U_i)$ **do**
14: **for all** $K_r \in T'(I_u)$ **do**
15: **if** $K_r \notin T'(I_v)$ **then**
16: $\mathbb{IK}.appendUnique(\mathbb{IK}_{K_p})$
17: **end if**
18: **end for**
19: **end for**
20: **end for**
21: **end for**
22: **end if**
23: **end for**
24: **return** $\mathbb{IK}$

The ranking values are determined by the procedures presented in section 5.2. Section 6.2.2, presents the results of this method.

5.1.2 Classification Analysis

The latest methodology development phase consists in training a classifier, which is able to decide if a keyword or location is relevant to the user.

The classifier determines whether an artifact is indirect or not. This process is applied to all possible artifacts, which results in a long list of candidate artifacts. Ultimately, this list should be ranked, and a subset of the list is used for recommendation. Summarizing, the classification approach involves the following steps:

- List all possible artifacts;
- Classify each artifact as indirect or not;
- Remove all artifacts that are classified as not indirect;

- Sort the list according to a ranking strategy;
- Determine a subset of artifacts according to a threshold.

The classifier model used is the Support Vector Machine – SVM (see section 2.4.2). In our experiments, we try to decide if an artifact is relevant to a user, therefore, to select the classifier training features we identified those which represent relationships in the data, such as:

- The number of users of the artifact;
- The number of items of the artifact;
- The number of keywords which share users with the artifact;
- The number of user who use the artifact and other artifacts of the user;
- The number of keyword which co-occur with each artifact of the user.

Sections 6.1.0.2 (keywords) and 6.1.0.3 (locations) describe the features which enabled the best results.

5.2 Ranking Results

Sorting entries on an indirect artifacts' list requires ranking methods designed for each type of artifact.

5.2.1 Ranking Indirect Keywords

To sort the list of indirect keywords we propose a method, R_{K_p}, which for every keyword, K_p, with a positive match, is defined by:

$$R_{K_p} = \sum_{K_r \in UK(U_i)} |\{K_r : \exists I_t \in I : K_r \in T(I_t), K_p \in T(I_t)\}|$$

R_{K_p} calculates the sum of the number of co-occurrences between K_p and the user's keywords. We refined this method with the approach used by Sigurbjörnsson and Van Zwol [67] and normalized the ranking values by the frequency of K_p, F_{K_p}, resulting R'_{K_p} such that:

$$F_{K_p} = \frac{|\{I_t : K_p \in T(I_t)\}|}{|I|}$$

$$R'_{K_p} = \frac{R_{K_p}}{F_{K_p}}$$

5.2.2 Ranking Indirect Locations

To sort the list of indirect locations we propose a method, R_{L_l}, which for every location, L_l, with a positive match, is defined by:

$$R_{L_l} = \sum_{L_g \in UL(U_i)} |\{L_g : \exists I_t \in I : L_g \in G(I_t), L_l \in G(I_t)\}|$$

R_{L_l} calculates the sum of the number of co-occurrences between L_l and the user's locations.

Following the same approach as the indirect keyword ranking strategy, we refined the method to compute, R'_{L_l}, which is the normalization of R_{L_l} by the frequency of L_l, F_{L_l}, and is defined by:

$$F_{L_l} = \frac{|\{I_t \mid L_l \in G(I_t)\}|}{|I|}$$

$$R'_{L_l} = \frac{R_{L_l}}{F_{L_l}}$$

5.3 Network Sampling

To evaluate our methodology, we created graph samples replicating the network structure at a particular moment in time, which is different to samples which represent the same network only with fewer nodes and edges. The main difference between goals is that the former preserves the network growth properties, which is ideal to evaluate our methodology with networks at different growth stages.

We adapted the sampling method by Leskovec and Faloutsos. [30] (see section 2.4.1) to deal with the structure of the information present in productive networks, yielding graph samples which verify both properties we are interested in preserving. Algorithm 2 presents the sampling method.

Algorithm 2 Network graph sampling method, which preserves the relevant aspects of productive networks. Visiting a node implies sampling it.

```
1: Randomly select a seed node to visit.
2: while exists a node to visit do
3:     Visit one node.
4:     Decide, with α probability, if links should be visited.
5:     if links should be visited then
6:         All outgoing nodes are selected to visit.
7:     end if
8: end while
```

5.4 Metrics

To evaluate and determine the effectiveness of the classification analysis results, we adopted standard metrics. Our focus is also on the evaluation of ranked lists of results, which are evaluated with specific metrics.

These metrics measure the effectiveness with which information retrieval systems answer queries. In this context, a query is a request for indirect artifacts.

The classification analysis queries for *relevant* artifacts, which are part of a whole set of *retrieved* artifacts. All metrics are defined in terms of relevant and retrieved artifacts.

We present all metrics in the context of an example. Consider a set of objects which are assigned to one of two classes, A and B. We trained a classifier to assigned a class to new data items, and we want to retrieved all items with class A. Table 5.1 shows the results of the classification process, where 6 items are retrieved, 4 of which are actually relevant, out of 5 relevant items in the dataset.

Furthermore, consider that we deliver the results to the user ranked by item number.

Table 5.1: Classification results from an hypothetical classifier, which assigns classes A or B to data items.

Item	Actual Class	Predicted Class
1	A	A
2	B	A
3	B	B
4	A	A
5	A	A
6	B	B
7	A	B
8	B	B
9	A	A
10	B	A

5.4.1 Precision

Precision (P) is the proportion of retrieved artifacts that are relevant. It is determined by

$$P = \frac{|relevant \cap retrieved|}{|retrieved|} \tag{5.1}$$

In our example, $P = 6/4 = 0.667$. We conclude that 67% of the retrieved items are relevant.

5.4.1.1 Precision at Rank

Precision at a rank K ($P@K$) report the proportion of the top K retrieved artifacts that are relevant. Precision at a specific rank is relevant because only the top results are ultimately returned to the user. Given that our goal is to present a list of recommendations, which cannot be too long to be effectively delivered by most user interfaces, we show the precision at rank 1, 5, 10, and 20.

Table 5.2 shows the precision at different ranks in our example.

Table 5.2: Precision at several ranks for the example retrieval results.

P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10
1.000	0.500	0.500	0.667	0.750	0.750	0.750	0.750	0.800	0.667

5.4.2 Recall

Recall (R) is the proportion of relevant artifacts that are retrieved. It is determined by

$$R = \frac{|relevant \cap retrieved|}{|relevant|} \tag{5.2}$$

The main focus of our methods is not to retrieve all artifacts that are relevant to the user, but instead to ensure that the ones that are retrieved are indeed relevant. However, to access the quality of the approach we include the recall computation, which provides a measure of missed relevant results.

In our example, $R = 5/4 = 0.800$. We conclude that 80% of the relevant items are retrieved.

5.4.3 F_1 Score

The F_1 score represents the accuracy of a query, in a scale from 0 to 1. Values close to 1 indicate better performances. It is computed by the harmonic mean of precision and recall, as described by

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{5.3}$$

In our example, $F_1 = 2 \times \frac{0.667 \times 0.800}{0.667 + 0.800} = 0.727$.

5.4.4 Mean Reciprocal Rank

The Mean Reciprocal Rank (MRR) informs where the first relevant artifact occurs in the ranking. It is determined by

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i} \tag{5.4}$$

In our example we have 5 relevant items (i.e., with class A). Consider that we query the system for each of the relevant items, and the results, for each query, were lists where the item appears in a rank equal to its item number.

Our queries are able to find 4 of the 5 relevant items, with ranks 1, 4, 5 and 9, respectively, therefore,

$$MRR = \frac{1}{5} \times \left(\frac{1}{1} + \frac{1}{4} + \frac{1}{5} + \frac{1}{9} \right) = 0.312$$

5.5 Discussion

The methodology partially addressed the research questions concerned with the discovery of indirect relationships – **RQ 3**, **RQ 6**, and **RQ 4**. Chapter 6 presents the evaluation.

A fundamental aspect of the methodology is that it identifies the main challenge of the indirect relationships discovery: the identification of indirect artifacts, i.e., keywords or locations. From that, the methodology relies on the productive network model to describe its operations, and its description includes an evaluation strategy.

Chapter 6

Model and Methodology Evaluation

Summary

This chapter presents the evaluation of the indirect relationship discovery methodology using both types of indirect artifacts: *indirect keywords*, and *indirect locations*. We present experimental protocols designed to evaluate both strategies, achieving positive results, which are an improvement on our related work.

The evaluation addresses research question **RQ 6**, which asks for a measure of effectiveness of the indirect artifacts discovery method.

6.1 Experimental Protocol

The experimental protocol for all types of indirect artifact (i.e., keywords and locations) contexts is similar.

Considering indirect keywords, we propose that if a user used a keyword to annotate an item, then it would be a valid suggestion in a modified scenario where this annotation had not been performed by the user. The test consists of verifying if the method can, in a situation like this, propose the keyword as indirect to the user. Therefore, the experiment strategy is to remove one keyword from the set of direct keywords associated to the user's annotations. The experiment outputs a ranked list of indirect keywords, in which the removed keyword should appear. We refer to this procedure as keyword, or, generally, artifact, *recovery*.

Similarly, removing the relationship between user and direct location, and trying to identify – recover – it as indirect location shows that the experimental method is suitable for indirect location discovery.

Algorithm 3 presents the outline of the evaluation procedure for indirect artifact discovery, for a particular user, U_i. The methods *methodology* and *rank* must be defined according to the actual evaluation approach. Our research plan includes two methods to instantiate the *methodology* method: a frequency analysis approach; and a machine learning approach, using support vector classifiers. We begin with a frequency analysis to understand the dataset, which in turn will enable the identification of relevant features for the classification analysis.

Algorithm 3 Experiment outlines. According to context, presents the removal of the association between user and artifact, which the *methodology* is designed to recover. The *rank* function returns the position of an element in a list.

a) Indirect keyword discovery evaluation procedure

Require: $O(U_i) \neq \{\emptyset\}$
1: $T'(I_t) = list()$
2: **for all** $I_t \in O(U_i)$ **do**
3: **if** $T(I_t) \neq \{\emptyset\}$ **then**
4: **for all** $K_p \in T(I_t)$ **do**
5: **for all** $K_q \in T(I_t), K_q \neq K_p$ **do**
6: $T'(I_t).append(K_q)$
7: **end for**
8: $\mathbb{K}_r = methodology(T'(I_t))$
9: **print** $K_p \in \mathbb{K}_r$?
10: **print** $rank(K_p, \mathbb{K}_r)$
11: **end for**
12: **end if**
13: **end for**

b) Indirect location discovery evaluation procedure

Require: $O(U_i) \neq \{\emptyset\}$
1: $G'(I_t) = list()$
2: **for all** $I_t \in O(U_i)$ **do**
3: **if** $G(I_t) \neq \{\emptyset\}$ **then**
4: **for all** $L_l \in G(I_t)$ **do**
5: **for all** $L_m \in G(I_t), L_m \neq L_l$ **do**
6: $G'(I_t).append(L_m)$
7: **end for**
8: $\mathbb{L}_r = methodology(G'(I_t))$
9: **print** $L_l \in \mathbb{L}_r$?
10: **print** $rank(L_l, \mathbb{L}_r)$
11: **end for**
12: **end if**
13: **end for**

Given that the user explicitly annotated items with the removed artifact, we are sure that it is relevant to the user. Therefore, as presented in algorithm 3, for each artifact whose relationship with the user is removed, the experiment goal is to twofold:

1. Use the *methodology* method to identify it as indirect;
2. Attribute it a high rank value in the list of indirect artifacts.

With a set of indirect artifacts, we may define a user-to-user relationship recommendation, relying on a ranked list of users, as presented in algorithm 4.

6.1.0.1 Frequency Analysis

Our first approach was to analyze the datasets, and extract artifact characterization measures. The procedure output was:

1. The recovery result for each top keyword;

Algorithm 4 Outline of the procedure used to build lists of indirect relationships.

a) Potential relationships from indirect keywords.

Require: $O(U_i) \neq \{\emptyset\}$
result = list()
if $\mathbb{IK}(U_i) \neq \{\emptyset\}$ **then**
  for all $K_p \in \mathbb{IK}(U_i)$ **do**
    for all U_j *such that* $K_p \in \mathbb{IK}(U_j)$ **do**
      result.append(U_j)
    end for
  end for
end if
return rank(result)

b) Potential relationship from indirect locations.

Require: $O(U_i) \neq \{\emptyset\}$
result = list()
if $\mathbb{IL}(U_i) \neq \{\emptyset\}$ **then**
  for all $L_l \in \mathbb{IL}(U_i)$ **do**
    for all U_j *such that* $L_l \in \mathbb{IL}(U_j)$ **do**
      result.append(U_j)
    end for
  end for
end if
return rank(result)

2. The ranking score (see definition 9);

3. The general characterization of the keyword, i.e., the total number of items and direct keywords.

6.1.0.2 Support Vector Classifiers for Indirect Keyword Discovery

To represent the training features of a classifier we use the notation presented in section 2.4.2:

$$\mathcal{F} = \langle |\mathcal{F}_1|, ..., |\mathcal{F}_n| \rangle$$

For a keyword, K_p, removed from the user's (U_i) direct keywords, by the procedure in algorithm 3.a), we propose two pairs of feature sets, $\mathcal{F}^a$ and $\mathcal{F}^b$, defined by:

$\mathcal{F}^a$ Each keyword is represented by the number of users that use the keyword and other keywords of the user ($\mathcal{F}^a_1$), and the number of keywords that co-occur with it in the user's items ($\mathcal{F}^a_2$):

$$\mathcal{F}^a_1 = \{U_j \mid \forall K_q \in UK(U_i) : K_q \in UK(U_j)\}$$
$$\mathcal{F}^a_2 = \{K_q \mid \forall I_t \in O(U_i), K_q \in T(I_t), K_p \in T(I_t)\}$$

$\mathcal{F}^b$ Each keyword is represented by it's absolute number of items ($\mathcal{F}^b_1$) and it's absolute number of users ($\mathcal{F}^b_2$).

$$\mathcal{F}^b_1 = \{I_t \mid \forall I_t \in I : K_p \in T(I_t)\}$$
$$\mathcal{F}^b_2 = \{U_j \mid \forall U_j \in U : K_p \in UK(U_j)\}$$

6.1.0.3 Support Vector Classifiers for Indirect Location Discovery

For a location, L_l, of the user's (U_i) direct locations, we propose one pair of feature sets, $\mathcal{F}^c$, defined by:

$\mathcal{F}^c$ Each location is represented by its absolute number of items ($\mathcal{F}_1^c$) and its absolute number of users ($\mathcal{F}_2^c$).

$$\mathcal{F}_1^c = \{I_t \mid \forall I_t \in I : L_l \in G(I_t)\}$$
$$\mathcal{F}_2^c = \{U_j \mid \forall U_j \in U : L_l \in UL(U_j)\}$$

6.2 Flickr Experiment

Our first experiment uses data sampled from the Flickr network (see chapter 3). The goal is to evaluate the methodology for indirect keyword discovery.

6.2.1 Dataset Description

Flickr provides an API [1] that facilitates querying its content. Through the API, it is trivial to obtain a user characterization from the user name or id. It is also possible to obtain a user's list of photographs and one photograph's list of keywords. The API also allows the querying of the system for a particular keyword, providing, as a result, the list of photos associated with the keyword.

The sample graph (see section 2.4.1 for the sampling method) represents 912 users, 249 151 items and 116 662 keywords. It contains 2 698 127 edges between items and keywords.

6.2.2 Frequency Analysis Results

We performed the frequency analysis with the top 20 keywords of each user.

The first conclusion drawn from the results was that the ranking method through the number of direct keywords does not produce meaningful results, and we excluded it from further analysis. The remaining analysis is focused on the ranking through the number of items.

Table 6.1 shows the keyword frequency analysis over the number of items.

We consider two sets of keywords: all the keywords in the dataset, and the set of keywords that were removed, i.e., the top keywords of each user. We see that the average number of items for the keywords removed during the test case is 364.76. This value is relevant while analyzing the recovery results.

[1] The Flickr API can be accessed through urls in the form of "`http://api.flickr.com/services/rest/?method=`", and its documentation is available at `https://www.flickr.com/services/api`

Table 6.1: Characterization of the number of items associated with a keyword.

Keywords	Mean	Std. Dev.	Mode	Minimum	Maximum	Percentiles		
						25%	50%	75%
All	23.12	191.90	1	1	15513	1	1	5
Test Case	364.76	964.44	1	1	15513	8	67	264

The threshold values used to determine a recovery were estimated after an analysis of the average recovery rate of several thresholds, between 10 and 5000, with increments of 10. These results are available in Fig. 6.1.

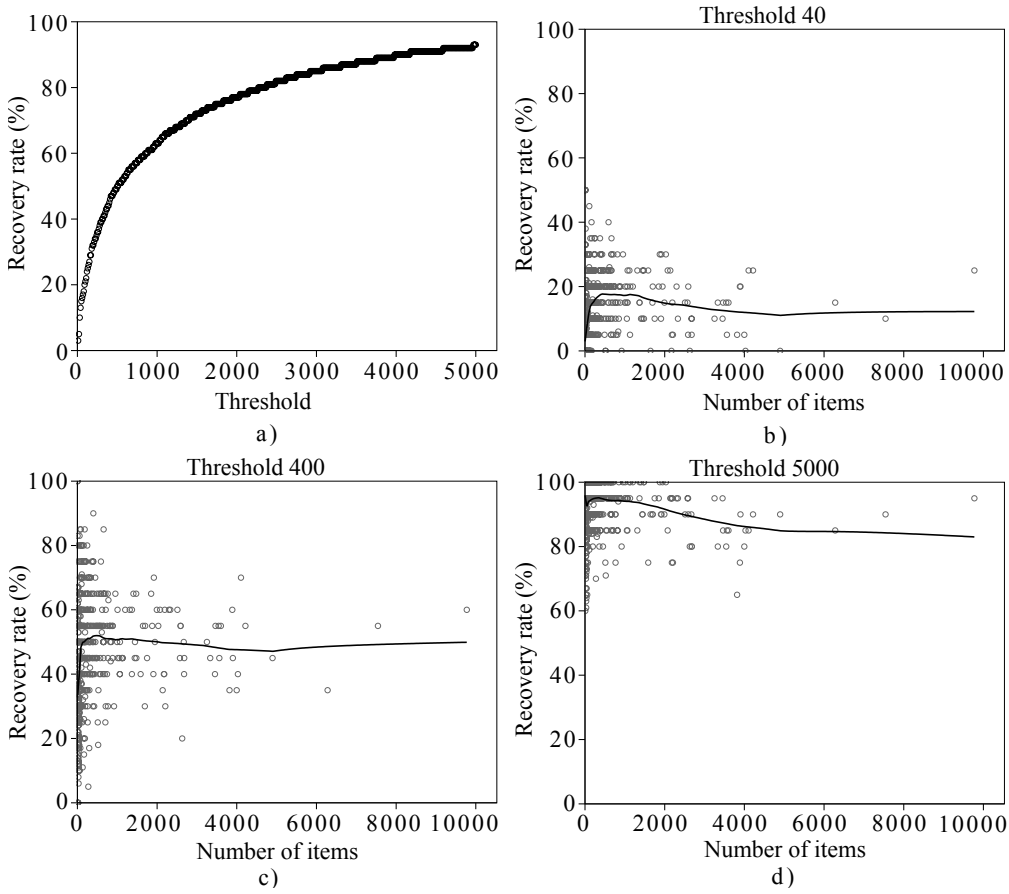

Figure 6.1: Frequency analysis results for the users' top 20 keywords' average recovery rates by the user's number of items: a) average recovery rate of thresholds between 10 and 5000, in intervals of 10; b) recovery rate for threshold 40; c) recovery rate for threshold 400; d) recovery rate for threshold 5000. Shows the users' top 20 keywords' average recovery rates by the user's number of items.

We now focus on three thresholds: one at 40 (shown in figure 6.1), which represents the minimum meaningful value, below which there are no useful recovery rates; one at 5000 that analyses the maximum extreme in Fig. 6.1; and one at 400 that explores values

around the average number of items per keyword in the test case (364.76). The average recovery rates (with SD standard deviation and MD mode) are: 12.74 (SD=10.71, MD=0) for threshold 40; 45 (SD=17.45 MD=50) for threshold 400; and 92.6 (SD=8.47, MD=50) for threshold 5000.

The keyword frequency characterization reveals that although the average number of items associated with a keyword is low, these are highly skewed towards much higher values. However, the mode is 1, which means that most keywords are associated with only one item.

As presented in figure 6.2, we found a significant correlation between the recovery rate and the number of items of the user (Spearman correlation coefficient of 0.81, p-value < 0.0005), which is consistent with the lower recovery rate for users with a high number of items, because keywords that are exclusive to the user cannot be recovered by our method – there are no paths in the graph that connect the keyword.

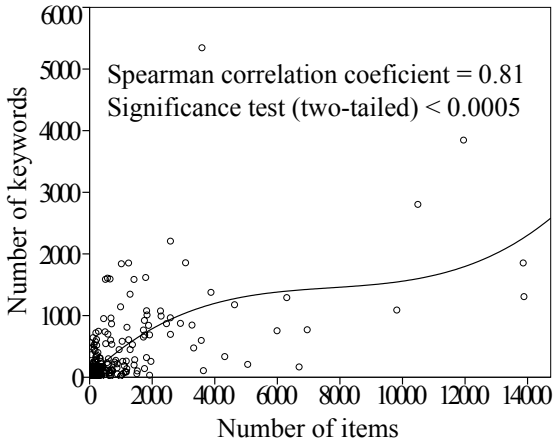

Figure 6.2: Correlation between unique and exclusive keywords and the number of items of a user.

6.2.3 Classification Analysis Results

The evaluation uses two set of users: a set of 50 users and a set of 300 users. In each we query for 50 keywords for each user. We show the the results for both training sets of features, i.e., $\mathcal{F}^a$ and $\mathcal{F}^b$, described in section 6.1.0.2.

Table 6.2: Classification task evaluation results.

Number of Queries	Rank	MRR	P@1	P@5	P@10	P@20
50	$\mathcal{F}^a$	0.5064	0.3600	0.3240	0.3080	0.2520
	$\mathcal{F}^b$	0.6372	0.5200	0.3840	0.3140	0.2580
300	$\mathcal{F}^a$	0.2817	0.1800	0.1180	0.0977	0.0783
	$\mathcal{F}^b$	0.3978	0.2667	0.2027	0.1690	0.1355

Table 6.2 shows the results ranked using the sum of the number of co-occurrences between each keyword and the user's keywords, normalized by the frequency of the keyword, i.e., R'_{K_p}, as described in section 5.2.

Although we were not able to reproduce results of keyword or user recommendation methods in the same context as ours, some authors present work that is comparable to ours. The main differences are the datasets, which unfortunately we were not able to obtain. In table 6.3 we partially reproduce the authors results, and compare them with our best method, $\mathcal{F}^b + R'_{K_p}$, where we obtain an improvement in the mean reciprocal rank over Chao Zhou et al. [84], and precision over Zhou et al. [42].

Table 6.3: Comparison with related work.

Method	MRR	P@1
$\mathcal{F}^b$	0.3978	0.2667
Chao Zhou et al. [84]	0.2345	0.3272
Zhou et al. [42]	-	0.33 [2]and 0.19 [3]

6.3 Twitter Experiments

In this section we present the results of a second set of experiments, which include indirect location discovery.

6.3.1 Dataset Description

The evaluation of the model and methodology uses 6 datasets built with Twitter data. Table 6.4 summarizes the datasets.

We designed a live Twitter feed capture tool that collects and organizes the information according to our information model. All datasets originated from a particular event that we were able to monitor (live music summer events, in Portugal).

The information collected contains users, items, keywords, locations, and places, where places are locations to which Twitter assigned some semantic (e.g., streets and businesses names). However, for the datasets available, the number of places is relatively low. We consistently obtained a very low percentage of geo-referenced information. Section 6.3.3.2 presents a discussion on the topic, and the impact of the low (or absent) number of located tweets on indirect location recommendation.

Table 6.5 describes the datasets, with counts of the several dimensions available.

[2]Dataset with 5 000 items [42].
[3]Dataset with 30 000 items [42].

Table 6.4: Twitter datasets available for evaluation. Each dataset was obtained by collecting the live feed resulting from filtering the Twitter stream with the given queries.

ID	Event Description	Query
E1	Rock in Rio Lisboa music festival	#rirlx
E2	Lisbon late spring holidays ("Santos Populares")	#santospopulares
E3	Lisbon Mega Picnic	#megapicnic
E4	Paredes de Coura music festival	#rirlx
E5	Paredes de Coura music festival V2	#paredesdecoura #vodafoneparedesdecoura
E6	Sol da Caparica music festival	#soldacaparica

Table 6.5: Description of the datasets. The number of *places* is indicated in the locations' column, in parenthesis.

ID	Items	Users	Keywords	Locations
E1	47114	26750	1820	743 (287)
E2	558	356	546	79 (14)
E3	16	14	5	2 (2)
E4	375	177	203	0 (7)
E5	908	325	365	0 (13)
E6	303	188	168	0 (6)

All datasets contain the complete set of tweets associated with the events, starting 72 hours before the event begins, and ending 72 hours after it closes. However, only the first two, E1, and E2, contain enough spatial data to enable indirect location identification.

6.3.2 The Need for Identifying Location Clusters

As described before, the datasets contain a low percentage of geo-referenced items (as expected). Moreover, the relationship pattern between item, keyword and location proved to be insufficient to enable our evaluation. The method requires a set of keywords to be associated with locations, through items. However, in most cases, each location correspond to only one item. Such is caused by the granularity operated by the GPS sensor on the mobile devices used to create the item. The same user, posting twice from the same location, a few minutes apart, is likely to produce two different coordinate pairs.

The solution to the problem is clustering locations. Instead of running the evaluation directly on locations, we compute a set of location clusters, using the DBSCAN [13] algorithm. Algorithm 5 presents the clustering procedure outline, including the information

cross-referencing computation between single locations and the respective clusters.

Algorithm 5 Procedure used to build location clusters and associate the information needed for the classification analysis with clusters (instead of single locations).

Require: $UL(U_i) \neq \emptyset$

1: $clusters = DBSCAN(UL(U_i), eps, min\_pts)$
   {*clusters* is a collection of location clusters.}
   {Each *cluster* contains a set of locations.}
2: **for all** *cluster c* **in** *clusters* **do**
3: $c.items = \{I_t \mid \forall L_l \in c.locations, G(L_l, I_t)\}$
4: $c.users = \{U_i \mid \forall I_t \in c.items, O(I_t, U_i)\}$
5: $c.keywords = \{K_p \mid \forall I_t \in c.items, T(K_p, I_t)\}$
6: **end for**

The choice of parameter values for DBSCAN was not focused on optimal behavior in terms of clustering. The problem lies with the high amount of locations, most with only one associated keyword, so we are mainly looking to significantly reduce the number of locations. The informal heuristic we followed was to obtain a number of clusters equal to around 10% of the number of locations in the datasets, merging items, users, and keyword sets of cluster members, thus producing clusters with more than 1 keyword on average. Figure 6.3 shows the clustering results (and parameters used) on the E1 dataset, reducing from 743 locations to 79 location clusters.

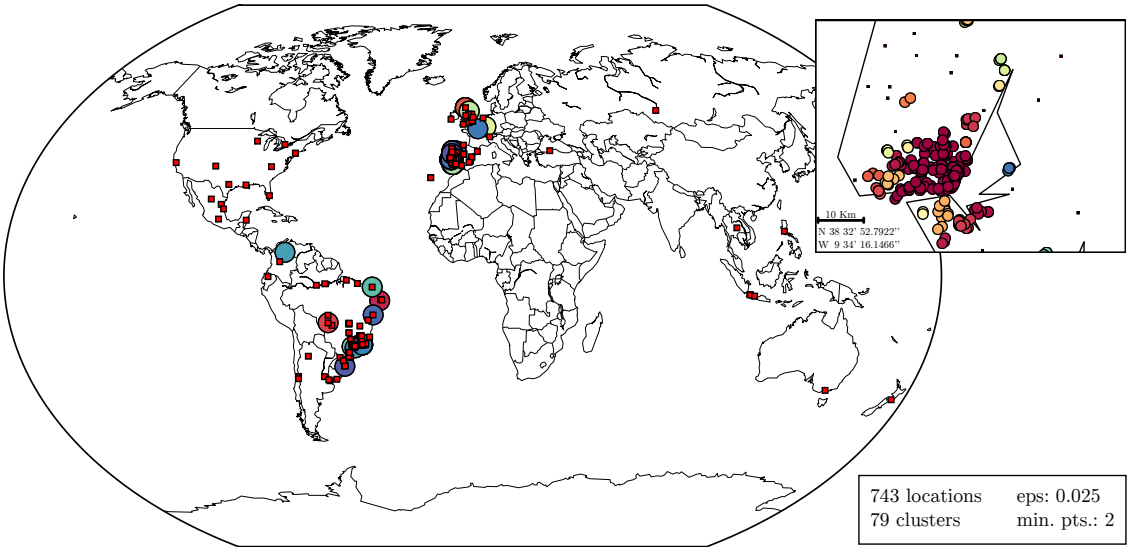

Figure 6.3: Clustering results for dataset E1. DBSCAN parameters are set to produce around 10% of the initial amount of locations. Circles represent locations in clusters, red squares represent noise. The world map (Kavrayskiy VII projection) shows all 79 clusters and noise, and the top right map (orthographic projection) shows results around Lisbon, Portugal (coordinates displayed for the lower left corner).

Section 6.3.3.2 discusses the classification results between the clustering approach and the original set of locations.

6.3.3 Classification Analysis Results

6.3.3.1 Indirect Keywords

Our first set of results are from the indirect keyword analysis. These are relevant to establish a comparison between our model and the original productive network model, and to determine if there is a substantial difference between the datasets extracted from Twitter, and the Flickr dataset used by the original model.

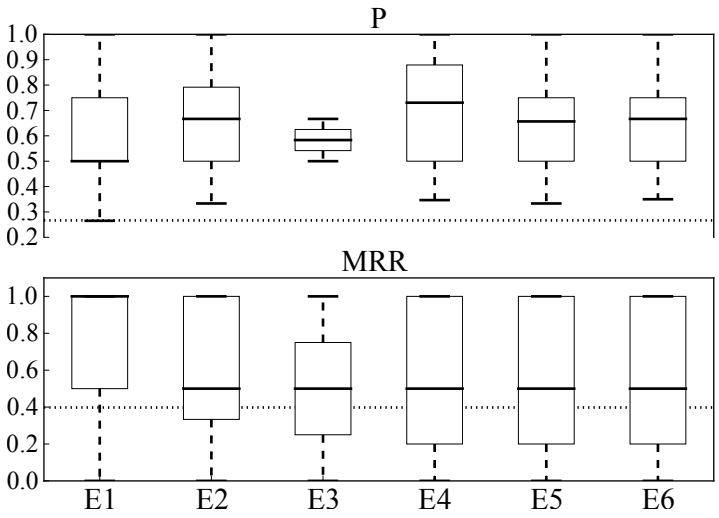

Figure 6.4: Boxplots with the Precision (P), and Mean Reciprocal Rank (MRR) results, for each case study. Both metrics are computed for each user of each case and this figure shows the average distribution of the metrics for each dataset. The horizontal doted lines represent the results obtained by the first set of experiments with Flickr (MRR=0.3978, P@1=0.2667 – see section 6.2.3).

Figure 6.4 shows the distribution of the results for the Mean Reciprocal Rank and Precision for the 6 datasets, contextualized by results of the Flickr experiment previously presented. There is a consistent difference on the Precision, which is explained by the comparatively smaller size of the datasets. The Mean Reciprocal Rank results are similar between experiments.

We conclude that the indirect keyword discovery methodology was successfully replicated. The better precision results are due to size differences between datasets which, in the Flickr case study, generate considerable noise.

6.3.3.2 Indirect Locations

With a successful replication of the indirect keyword discovery methodology we now focus on indirect location discovery with location clusters. Table 6.6 shows the results of method $\mathcal{F}^c$ (see section 6.1.0.3).

Given the low percentage of keyword-location association, only the first (E1) dataset allows results without the clustering approach. Clustering significantly improves the Mean Reciprocal Rank results, and allows indirect location discovery on smaller datasets.

Table 6.6: Results of the indirect locations classification analysis. E1 and E2 represent datasets with clustering. E1* is the original dataset, without clustering.

ID	MRR	P@1	R@1
E1	0.5390	0.6415	0.4351
E1*	0.3785	0.6259	0.4118
E2	0.1365	0.7371	0.5804

6.4 Discussion

The evaluation of the indirect relationship discovery methodology focused on the effectiveness of the methodology in finding indirect elements which enabled indirect relationships. Indirect elements defined links between users in the productive network graph, and each user has a ranked list of indirect elements. These elements are keywords and locations.

Indirect keywords and locations of one user are used to annotate the content of several other users, which make them candidates for relationships.

The evaluation is a complement of the methodology presented in chapter 5, addressing the research questions **RQ 3** and **RQ 6**, focused on the identification of indirect elements.

The methodology evaluation also addresses research question **RQ 4** by showing that we may identify potentially relevant relationships. This potential is supported by the ability to discover keywords and locations of interest to the user.

Finally, the evaluation addresses **RQ 7**, which asks for a measure of effectiveness of the methodology.

Chapter 7

Applications

Summary

This chapter presents applications focused on relationships in productive networks. These applications are enabled by the productive network model, presented in chapter 4, and the indirect relationship discovery methodology, presented in chapter 5. We present a visualization and interaction platform to explore and provide context to relationship in the networks, and a preliminary work on population density estimation, based on indirect relationship between locations and users.

7.1 Indirect Relationship Visualization Platform

From a user's perspective, the complex network relationships of productive networks, enabled by the graphs presented in section 4.1 (specifically, table 4.1 and table 4.2) is not easily grasped. To provide insight on the relationship network, we developed a web-based graph visualization and interaction platform, which enables the visual identification of relationships in a productive network.

Given that the platform's goal is to provide awareness about relationships on the network, it should be able to inform about the context of that relationship and avoid visual cluttering, by selectively show and hide information. The user should be able to navigate through relationships, which may involve changes in the context and require real time decisions to avoid cluttering. Table 7.1 presents the design goals for the platform development.

The platform is based on a graph structure, with nodes representing information elements, and edges as connections between those elements. The information model supports 12 different types of graphs, each with a different combination of basic concepts

as nodes and edges. However, not all combinations are useful for a particular network. Therefore, catering for our particular case study (see section 7.1.1), the platform enables 3 different perspectives over the network, each tailored to visualize different contexts: commonly annotated items; users who share keywords; or keywords associated with items. The platform is extensible to support the remaining graph representations.

Table 7.1: Design goals for the interaction and visualization platform.

Design Goal	Description
DG1	Selectively display the relevant information to understand a particular set of relationships.
DG2	Provide awareness about the context of any given visualization.
DG3	Enable changes in the focus of the visualization, between users, items, and keywords.
DG4	Avoid visual clutter.
DG5	Enable navigation.

The platform enables the construction of tools that provide insight over the network, increasing awareness for the potential of indirect relationship discovery. The platform and its interactive capacities have been evaluated through the implementation of a case study, based on a sample of the Flickr[1] network. The case study was evaluated through user studies, with positive results.

To build one of the graphs in table 4.1, we begin with a random or user selected node. The content of this initial selection restricts the type of content available for the node set. The type of edge is selected by the user.

The set of edges contains all the relevant edges for each node in the set of nodes, according to the specifications in table 4.1. Each element in the set of nodes is queried in the network to obtain all its relevant edges. When an element that is not in the set of nodes would enable an edge with an existing node, this element is included as a node, and is then queried in the network for edges.

This approach will eventually include all the nodes and edges in the network, leading to the visual clutter we are set to avoid.

To deal with the clutter, we propose two special types of node and edge, in order to represent sets instead of single elements: the *super node* and *super vertex*.

We select every initial element as a single node. Furthermore, each element that uniquely uses an edge to connect to a single node, is also a single node. All other elements are grouped in a super node. Similarly, each edge that connects to an edge (single or super), is represented as a single edge if there are no more edges that could connect the nodes. Otherwise, it is represented as part of a super edge. Table 7.2 presents further descriptions of single and super nodes and edges.

[1] http://www.flickr.com

Table 7.2: Operations available to the different element types

Element	Operations	
Single Node	*Selection*	*Expansion*
	Access to the node's information.	Access to nodes related through a chosen edge. If the results include more than one node, the nodes are grouped in a super node.
Super Node	*Selection*	*Extraction*
	Access to every single node that belongs to the super node.	Moves interior node from the group into the graph area. If every node is extracted, the super node disappears.
Single Edge	*Selection*	*Addition of a single node*
	Access to the edge's information.	Chooses one related node and adds it to the graph as a single node.
Super Edge	*Selection*	*Expansion*
	Access to every single edge contained in the super edge.	Moves nodes reachable from an interior edge of the group into the graph area. If the results include more than one node, these are grouped in a super node.

The representation of the graph is based on visual elements, which are identified by labels. Each label may be a name, an image, or both. It also uses the notion of super elements, which are collections of single elements, to deal with visual cluttering.

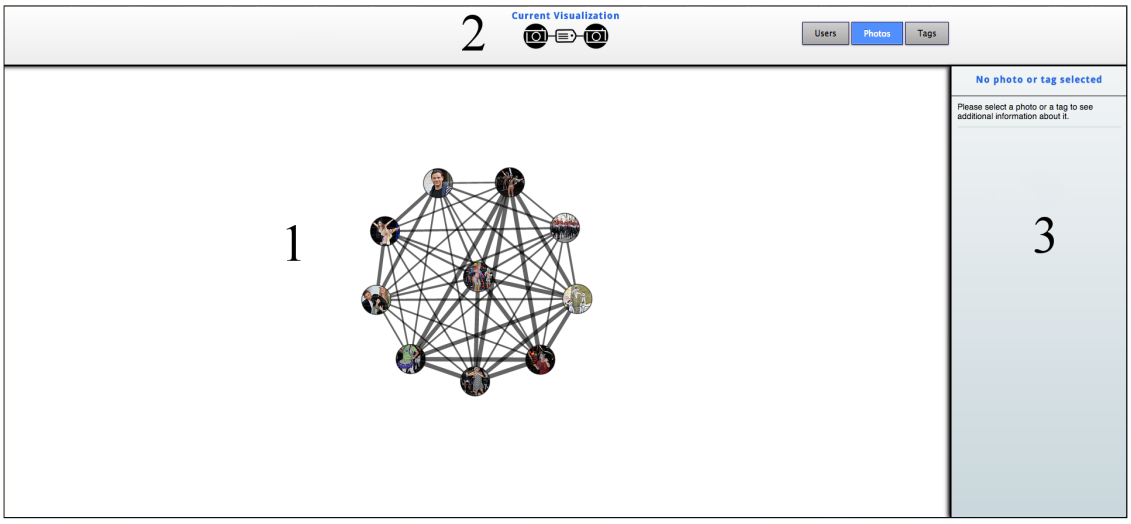

Figure 7.1: Visualization and interaction platform web-based prototype main area: 1. Graph area; 2. Top panel; 3. Right panel.

The platform enables visualizations using any combination between two of the productive network concepts (users, items, keywords, and locations) as a relationship. An example of a visualization is items, e.g., photos (as nodes), connected through common

keywords (as edges).

The web-interface prototype used for the evaluation does not include locations, and therefore does not enable graphs presented in table 4.2.

The web-interface has three main areas, presented in figure 7.1. The central area displays the interactive graph. On the top panel, the user may choose the type of visualization desired, and verify which is the current one. Finally, the right panel provides access to all the information relative to the currently selected graph element. All areas are interconnected and interaction in one may trigger a change in another (e.g., the selection of a graph element triggers a change of information on the right panel).

Figure **??** presents the visualization elements of all types of node and edges, and the interaction context on the right bar.

7.1.1 Case Study

The example used for the development and testing of the platform was Flickr, a Yahoo platform where users may share photos. Flickr's data model is easily represented to the productive network model: Flickr users, photos and tags are modeled into the model of users, items and keywords, respectively. In Flickr, users submit their photos and associate tags with them, thus also becoming individually associated with the tags. These associations enable several perspectives over the network, possible to be visualized in the platform: users connected by tags, photos connected by tags and tags connected by photos.

7.1.2 Evaluation

The platform was evaluated by 15 users, which submitted answers to two types of questionnaire. Section B.1, in appendix B, presents a detailed characterization of the users.

The first questionnaire was divided into two parts: the first part consists of a general appreciation, based on pair-valued adjectives; the second part addresses different features in the platform, focusing on usefulness and easiness of use.

The second questionnaire was made on Attrakdiff[2], also focused on general appreciation, using pair-valued adjectives, whose output is a mean value over two dimensions: *Pragmatism*, i.e., how it addresses its main purpose; and *Hedonism*, i.e., how it addresses the user experience.

Attrakdiff produces a score which represents the overall assessment of the interface, indicating user satisfaction and potential for improvement. However Attrakdiff does not make the questionnaire results available for further analysis, which prevents the performance assessment of individual visualization and interaction components. Such led to the decision to implement a general appreciation part in the first questionnaire. Furthermore, we present our questionnaire's design and result analysis in a format which may be replicated by similar studies.

[2]http://attrakdiff.de/sience-en.html

7.1.2.1 First Questionnaire: Design and Results

For the first questionnaire, the users were briefed about the study procedure, and were given a few minutes to explore the tool, without any instructions or guidance. Afterwards, there was a supervised trial, where the users were asked to use the tool to answer the scripted questions presented in section B.2, appendix B. Figure B.1, presents the answer times for all users.

Figure 7.2 presents the general appreciation results, which corresponds to the answers to question *Q1* of the questionnaire (presented in table B.1, appendix B). The outcome is positive, with most users considering the platform intuitive, pleasant, creative, useful, captivating and clear. Each answers is captured by a 5 level semantic differential scale. To avoid routine responses, the presentation of the pairs of opposite adjectives was designed to ensure that the value of 5 was not always associated with the most positive word. The radial plot in figure 7.2 shows normalized results.

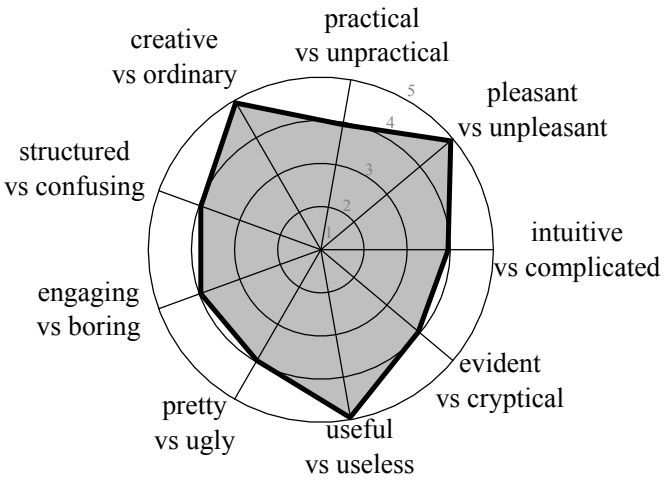

Figure 7.2: Summary of the results on the general appreciation of the platform. Scales of opposite adjectives were codified with values from 1 to 5. Normalization ensures that highest values are associated with the positive adjective, presented first in each label. Values show the median value of each pair of adjectives.

Table 7.3 summarizes the feature related results of the second part of the questionnaire (see table B.1, in appendix B). It shows query by navigation was successful in enabling the exploration of new information, which partially validates our goal of providing awareness about the context and confirms the navigation as a well suited interaction approach to switch between different network visualizations.

Users reacted positively to the amount of data that is generated (i.e., a single node or a super node) and understood the importance of hiding unwanted information and that the platform was successful in implementing this aspect. We successfully achieved our goal of avoiding visual clutter, while assuring that the user remains aware about the current visualization context.

Showing two components connected to each other is not always enough, and users felt

the need to switch to other types of connections under the context of a specific element to learn more about the data, which validates the goal of enabling changes in the focus of the visualization. However, the users pointed out that the switch button was not immediately seen and could be better highlighted.

Some users had difficulties in identifying the navigation context, which could be related with inconsistencies in the Flickr data, as presented in section 7.1.2.2.

Table 7.3: Platform features evaluation results, with references to questions of the questionnaire in parenthesis. Table B.1, in appendix B, presents the full questionnaire text and answer scales.

Dimension	Evaluation: [ordinal] - count (percentage)				
Usefulness in expanding a keyword (*Q2*).	Not Useful [1] 0 (0%)	[2] 1 (7%)	[3] 1 (7%)	[4] **8 (53%)**	Very Useful [5] 5 (33%)
Usefulness of the clustering of nodes (*Q3*).	Not Useful [1] 0 (0%)	[2] 0 (0%)	[3] 0 (0%)	[4] 3 (20%)	Very Useful [5] **12 (80%)**
Usefulness of the visualization switch (*Q4*).	Not Useful [1] 0 (0%)	[2] 0 (0%)	[3] 0 (0%)	[4] 5 (33%)	Very Useful [5] **10 (66%)**
Easiness in identifying the current context (*Q5*).	Hard [1] 0 (0%)	[2] 1 (7%)	[3] 2(13%)	[4] **8 (53%)**	Easy [5] 4 (26%)
Easiness in navigating to the desired context (*Q6*).	Hard [1] 0 (0%)	[2] 1 (7%)	[3] 1 (7%)	[4] **8 (53%)**	Easy [5] 5 (33%)

7.1.2.2 Attrakdiff Questionnaire: Results

The Attrakdiff questionnaire is a usability evaluation tool, which produces a value over two dimensions: the *pragmatic* qualities, which evaluates the behavioral consequences of the interface towards its objective; and the *hedonic* qualities, which assesses the emotional response of the user.

The results of the Attrakdiff questionnaire are summarized in figure 7.3. The user interface was rated as desired. In terms of hedonic quality, the results indicated that the users identified with the product and were motivated and stimulated by it. The confidence interval of the score indicates that there is little room for improvements.

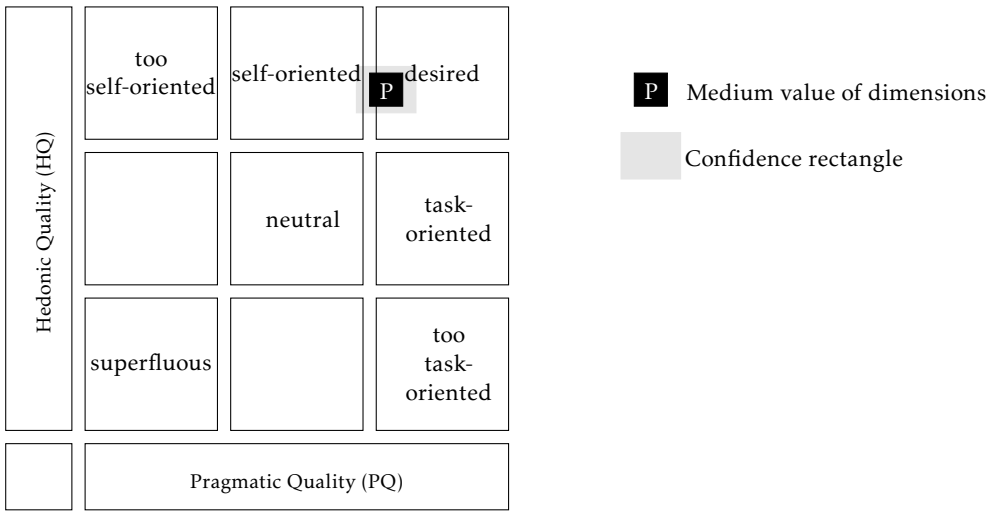

Figure 7.3: Attrakdiff results.

7.1.2.3 Case Study Limitations, Evaluation Threats and Constraints

This work focused on a Flickr dataset, which is also one of the datasets used for model evaluation (see chapter 6). Flickr does not enable users to share authorship photos, which limits the study, since it is not possible to identify current relationships with items (photos) as a connection, on this particular dataset. Flickr does not supervise the quality of the annotation process, leading to duplicated keywords and empty values.

Two different questionnaires were submitted to the users, with a time period between them, and both produced similar results. However, we acknowledge that the evaluation could gain from using two different groups of users, each answering one test.

The questionnaires answer value scales are semantic differential scales, which are analyzed as ordinal scales, because it is the less restraining assumption which does not threat their validity. This is the rational for using the median as a central tendency measure on the radial plot in figure 7.2.

7.1.3 Discussion

The general assessment of the platform is positive on both evaluation questionnaires. Moreover, the evaluation process produced several results which indicate that the framework successfully enables users to understand and explore productive network relationships. Table 7.4 presents a cross reference between the design goals, proposed in table 7.1, and the framework elements and evaluation results of the first questionnaire.

This work is presented in the following publications:

[63] [A. Sabino, J. Gouveia, and A. Rodrigues, "Visualizing productive networks", IADIS Int. J. WWW/Internet, vol. 12, no. 2, pp. 34–50, 2015.

Table 7.4: Cross reference between framework design goals (presented in table 7.1), its concepts, and evaluation results of the first questionnaire (referenced by question number, partially summarized in table 7.3, and described in detail by table B.1, section B.3, appendix B.

Design Goal	Framework Concept	First Questionnaire Questions
DG1	The node and edge selection operations.	Q3 (Very useful)
DG2	The top and right panels.	Q5 (Easy)
DG3	The top panel.	Q5 and Q6 (Easy)
DG4	Super nodes and edges.	Q3 and Q4 (Very useful)
DG5	The node and edge extraction operations, and the right panel.	Q2, Q3, Q4, and Q6 (Useful and easy)

[20] J. Gouveia, A. Sabino, and A. Rodrigues, “Visualizing productive networks relationships”, in Proceedings of the 13th International Conference WWW/INTERNET, 2014.

[60] A. Sabino, J. Gouveia, and A. Rodrigues, “Visualizing Productive Network Relationships”, in Proceedings of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence, 2014.

[21] Gouveia, J., "VisualAUThor - Uma plataforma de visualização de relações de colaboração", MSc Thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2013.

7.2 Population Density Estimation for Emergency Management

This section presents the outline of a research effort to address emergency management issues using the productive network model. We describe the methodology which is being implemented, and some preliminary results.

Assessing the risk of potential emergency situations caused by natural phenomena requires the study of the consequences of several scenarios on the affected areas. This risk assessment mainly considers the population density of the area, its land use, and natural environment.

The set of tasks involved in the risk assessment methodology vary according to the specific study area, but the risk calculation method is generally determined by the following [61]:

$$\text{Risk} = \text{Probability of occurrence} \times \text{Consequences} \tag{7.1}$$

In equation 7.1, two parcels characterize the risk associated to a crisis event: the frequency of the event; and the consequences associated with the event. The probability of an event is determined by the study of historical data and prediction models. The consequences are determined by an extensive analysis of data related with social, economical and environmental activity in the area.

Most systems perform an offline, static consequence analysis and produce a table of reference [16]. In this analysis, census data is the main source of information to estimate the population density of the area. However, although infrastructures and the local natural environment do not change in a short time interval, the human population density of the affected area may significantly vary, not only throughout the day, but weekly and seasonally.

Presently, with widespread access to the Internet, and mass adoption of mobile devices, the use of social networks services where the most widely used equipment of access was the smartphone, like Twitter[3] and Instagram[4], is becoming increasingly popular [36, 68]. The ubiquity inherent to these types of mobile devices, and the availability of sophisticated sensors (e.g. GPS), increases the precision and detail of the content of social networks.

We propose to develop a methodology to build dynamic social network user density maps for specific areas, using data from the networks. This is a step towards the dynamic mapping of population density, which would increase risk assessment accuracy by minimizing the difference between population estimation and the actual density.

Dynamic population density maps are particularly useful for early warning systems, where alerts are issued between a few days and a few hours before the event [15]. Our methodology is focused on the analysis of data for short time intervals, and relies on machine learning methods to relate users with potential future locations [57].

7.2.1 Preliminary Studies on Social Network User Density Variation Estimation

We present two preliminary studies over the Instagram network to evaluate the precision of its geolocated content. The two events were the Reef Hawaiian Pro Surf 2014 final in Hawaii, which occurred between 12 and 23 of November 2014, and one stage of World Rally Championship 2014 in Wales, which occurred between 13 and 16 of November 2014.

The Reef Hawaiian Pro Surf 2014 event area is not populated most of the year. However, while it never gets actually crowded, it attracts people interested in surf and beach activities during the summer. We focus our study in an event that occurs during the time

[3] https://www.twitter.com
[4] https://www.instagram.com

of the year when it is most populated, and try to detect the impact on population density during the surf championship event.

The World Rally Championship 2014, Wales stage event area is mostly unpopulated, and only attracts a few people interested in the landscape or water sports during the year. We focus on this area to evaluate the actual posts, and verify the content to determine if they are actually related with the event.

Table 7.5 presents the parameters used for data capture of both case studies, and the results.

Table 7.5: Instagram API query parameters and results for Reef Hawaiian Pro Surf 2014, and World Rally Championship 2014 Wales Stage.

Dataset	Collection Time	Coordinate		Radius	Results
		Lat.	Lon.		
Reef Hawaiian Pro Surf 2014	15/11/2014 (7h) to 16/11/2014 (23h)	21.5929º	-158.1088º	1 km	25 users
	18/11/2014 (7h) to 19/11/2014 (23h)	21.5929º	-158.1088º	1 km	3 users
WRC 2014 Wales Stage	15/11/2014 (7h) to 16/11/2014 (23h)	53.0812º	-3.5595º	5 km	98 posts
	20/11/2014 (7h) to 21/11/2014 (23h)	53.0812º	-3.5595º	5 km	20 posts

Each study is initiated by defining the request string to submit to the Instagram API. These are a geographic coordinate point and a radius. Requests are executed during and after the event, storing results in XML and JSON files for analysis, and to generate visual layouts with the results (with Google Maps[5] overlays), which enable response validation, and the comparison of the two time intervals.

Figure **??** presents a map with the results for both case studies. In the first study we requested users that have posted content about the event. Each data point, i.e., a user location, represents the location associated with the first post of that user. In the second study we collected all posts, therefore we expected to have more results with this approach, even if representing less people. While the goal is to count people, the second study enables an empirical evaluation of the post content, and verify that it is in fact related with the event.

7.2.1.1 Discussion and Future Work

Results show that, for the Reef Hawaiian Pro Surf 2014 event there were more users during the event (25), when compared with a regular day (3). Results hint that on event day there was a peak of population at the geographical area.

[5]https://maps.google.com

For the World Rally Championship stage, there were more posts during the event, when compared with the following day. We manually classified the posts in order to determine that they were actually related with the event. Results show that only one post had unrelated content. We conclude that the variation in the number of posts is caused by the event.

Results enable the conclusion that there are enough geo-referenced data on Instagram to map and detect changes in user density at a place, when capturing posts related with an event.

This work is presented in the following publication:

[50] J. Rosa, A. Sabino, and A. Rodrigues, "Monitoring social network user density variations in areas of interest", in Proceedings of the 18th AGILE International Conference on Geographic Information Science, 2015.

Future work will focus on the evaluation of the methodology as a tool to identify social network user density variations, which is a step towards population density variation estimation. Currently, this research effort deployed a tool to build datasets for this context.

Chapter 8

Conclusions

Summary

This chapter presents the final conclusions of our work, cross referencing out research questions with the contributions. We successfully studied a large collection of networks, enabling the proposal of a generic productive network model, which supports an indirect relationship discovery methodology, addressing our main research question.

This thesis presented studies, models and methods which enable the identification of indirect relationships on productive networks. Our background on emergency management set the challenge of developing methods for relationship discovery, to support tools such as expert finding, content recommendation and even population density variation estimation.

Motivated by that background, and in the context of an extensive literature review, we designed a network survey which provided evidence of a common set of information elements and operations in a type of networks we named *productive networks*.

We presented a model for productive networks, which enables a systematic representation of relationships between different network users and artifacts. This model, which naturally represents networks as graphs, is the main requirement for a machine learning based relationship discovery methodology we presented and successfully evaluated.

Our main research question was presented such as: *How to identify indirect relationships on social networks, content sharing networks, and content indexing networks?* Table 8.1 summarizes the questions which followed from our main research question, and ultimately guided our research.

Table 8.1: Thesis research questions.

Research Question	Description
RQ 1.	What are the common information concepts between social networks, content sharing networks and content indexing networks?
RQ 2.	How may we characterize productive networks in terms of the relationships between their underlying information concepts?
RQ 3.	May we identify relevant indirect topics of interest for productive network users?
RQ 4.	May we identify potentially relevant indirect relationships between productive network users?
RQ 5.	May we specialize the model to account for location based annotation systems?
RQ 6.	What is the effectiveness of indirect user and content discovery tools based on the productive network model?

8.1 Main Findings

The research question were addressed by several developments, from the productive network survey, model, and indirect relationship discovery methodology. Table 8.2 summarizes the developments that addressed each research questions.

Table 8.2: Developments which addressed each research questions, referencing the respective chapter.

Research Question	Development	Chapter
RQ 1.	The productive network survey.	Chapter 3
RQ 2.	The productive network model.	Chapter 4
RQ 3.	The indirect relationship discovery methodology.	Chapter 5
RQ 4.	The indirect relationship discovery methodology.	Chapter 5
RQ 5.	The productive network model and the indirect relationship discovery methodology.	Chapters 4 and 5
RQ 6.	The productive network model and indirect relationship discovery methodology evaluation.	Chapter 6

With an evidence supported productive network model, and a methodology for indirect relationship discovery with positive evaluation results, we have successfully addressed all research questions.

Ultimately, we *identify indirect relationships on social networks, content sharing networks, and content indexing networks* through machine learning methods, such as support vector

machines, in the framework of our productive network model. This approach is not only successful, but also valid for any network which may be represented as a productive network.

8.2 Applications and Future Directions

Some of our developments have already been transferred into applications. Chapter 7 presents a visualization and interaction tool designed to help users understand and discover information on productive networks.

Chapter 7 also presents an ongoing work towards population density variation estimation and forecast. So far the productive network model enabled results which correlate user density with geo-referenced item density in the same area. Future work will focus on the detection of users who are indirectly related with locations and on the forecast of population density variations, with a potential application in emergency management.

Finally, we evaluated the indirect relationship discovery methodology with experimental protocols designed to work with ground-truth data from our samples. This setup is optimal for the evaluation of content recommendation, but kept us from evaluating the recommendation of relationships between users. We plan to design a research project to integrate our proposals with a recommender system in a working productive network, and to conduct user trials.

Bibliography

[1] S. Apreleva and A. Cantarero. "Predicting the location of users on Twitter from low density graphs". In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015.

[2] M. Bakillah, R.-Y. Li, and S. H. Liang. "Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan". In: *International Journal of Geographical Information Science* 29 (2015).

[3] J. Bao, D. Lian, F. Zhang, and N. J. Yuan. "Geo-social media data analytic for user modeling and location-based services". In: *SIGSPATIAL Special* 7 (2016).

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Vol. 53. 2013.

[5] C. Buntain and J. Golbeck. "Identifying social roles in reddit using network structure". In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. New York, New York, USA: ACM Press, 2014.

[6] C. Fortes, R. Reis, M. Reis, P. Poseiro, R. Capitão, L. Pinheiro, J. Craveiro, J. A. Santos, S. Silva, J. Ferreira, M. Martinho, A. Sabino, A. Rodrigues, P. Raposeiro, C. Silva, A. Simões, E. Azevedo, F. Vieira, and M. Rodrigues. "Aplicação do Sistema HIDRALERTA na Avaliação do Risco Associado ao Galgamento no Porto da Praia da Vitória". In: *Actas do 3º Congresso Internacional de Riscos*. Guimarães, Portugal, 2014.

[7] S. Cetintas, M. Rogati, L. Si, and Y. Fang. "Identifying similar people in professional social networks with discriminative probabilistic models". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*. New York, New York, USA: ACM Press, 2011.

[8] E. H. E. H. Chi and T. Mytkowicz. "Understanding the efficiency of social tagging systems using information theory". In: *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia* (2008).

[9] N Cristianini and J Shawe-Taylor. *An introduction to Support Vector Machines*. Vol. 47. 2000.

[10] J. D. Cruz, C. Bothorel, and F. Poulet. "Community detection and visualization in social networks". In: *ACM Transactions on Intelligent Systems and Technology* 5 (2013).

[11] C. V. Damme, M. Hepp, and K. Siorpaes. "Folksontology: An integrated approach for turning folksonomies into ontologies". In: *Bridging the Gep between Semantic Web and Web 2.0 SemNet* (2007).

[12] A. F. Eisenberg and J. Houser. "Social Network Theory". In: *Encyclopedia of Sociology*. Ed. by G. Ritzer. 2007.

[13] M. Ester, H. Kriegel, J Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*. 1996.

[14] J. Ferreira, C. Fortes, M. Reis, P. Poseiro, A. Sabino, A. Rodrigues, S. Silva, J. Santos, R. Capitão, L. Pinheiro, J. Craveiro, P. Raposeiro, A. Simoes, E. Azevedo, M. Rodrigues, and C. Silva. "Sistema de Previsão e Alerta de Inundações em Zonas Costeiras e Portuárias – O Projeto Hidralerta". In: *XVI Encontro da Rede de Estudos Ambientais em Países de Língua Portuguesa e III Seminário Internacional de Ciências do Ambiente e Sustentabilidade*. Manaus, Brazil, 2014.

[15] C. Fortes, M. Reis, P. Poseiro, R. Capitão, J. Santos, L. Pinheiro, J. Craveiro, A. Rodrigues, A. Sabino, S. F. Silva, J. Ferreira, P. Raposeiro, C. Silva, M. Rodrigues, A. Simões, E. Azevedo, and F. Reis. "HIDRALERTA Project – A Flood Forecast and Alert System in Coastal and Port Areas". In: *Proceddings of the IWA World Water Congress and Exhibition*. Lisbon, Portugal, 2014.

[16] C. Fortes, M. Reis, P. Poseiro, R. Capitão, J. Santos, L. Pinheiro, A. Rodrigues, A. Sabino, M. Rodrigues, P. Raposeiro, J. Ferreira, C. Silva, A. Simões, and E. Azevedo. "O Projeto HIDRALERTA – Sistema de previsão e alerta de inundações em zonas costeiras e portuárias". In: *Proceedings of the 8th Jornadas Portuguesas de Engenharia Costeira e Portuária*. Lisbon, Portugal, 2014.

[17] C. J. Fortes et al. "Ferramenta de apoio à gestão costeira e portuária: o sistema hidralerta". In: *Proceedings of the VIII Congresso sobre Planeamento e Gestão das Zonas Costeiras dos Países de Expressão Portuguesa*. 2015.

[18] N. Garg and I. Weber. "Personalized, interactive tag suggestion for flickr". In: *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08* (2008).

[19] M. Gomez Rodriguez and M. Rogati. "Bridging offline and online social graph dynamics". In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. New York, New York, USA: ACM Press, 2012.

[20] J. Gouveia, A. Sabino, and A. Rodrigues. "Visualizing productive networks relationships". In: *Proceedings of the13th International Conference WWW/INTERNET*. Porto, Portugal, 2014.

[21] J. T. R. D. Gouveia. "VisualAUThor - Uma plataforma de visualização de relações de colaboração". MSc. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2013.

[22] L. Guo, J. Shao, K. L. Tan, and Y. Yang. "WhereToGo: Personalized Travel Recommendation for Individuals and Groups". In: *2014 IEEE 15th International Conference on Mobile Data Management*. IEEE, 2014.

[23] C. Hauff. "A study on the accuracy of Flickr's geotag data". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*. New York, New York, USA: ACM Press, 2013.

[24] M. Heckner, T Neubauer, and C. Wolff. "Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types". In: *Proceedings of the 2008 ACM workshop on Search in social media* (2008).

[25] V. Hegde, A. Mileo, and A. Pozdnoukhov. "Events Describe Places - Tagging Places with Event Based Social Network Data". In: *Proceedings of the 3rd IKDD Conference on Data Science, 2016 - CODS '16*. New York, New York, USA: ACM Press, 2016.

[26] P. Kefalas, P. Symeonidis, and Y. Manolopoulos. "A Graph-Based Taxonomy of Recommendation Algorithms and Systems in LBSNs". In: *IEEE Transactions on Knowledge and Data Engineering* 28 (2016).

[27] O. V. Laere, S. Schockaert, and B. Dhoedt. "Towards automated georeferencing of flickr photos". In: *Proceedings of the 6th Workshop on Geographic Information Retrieval*. 2010.

[28] T. Lappas and D. Gunopulos. "Interactive recommendations in social endorsement networks". In: *Proceedings of the fourth ACM conference on Recommender systems - RecSys '10*. New York, New York, USA: ACM Press, 2010.

[29] T. Lappas, K. Punera, and T. Sarlos. "Mining tags using social endorsement networks". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11* (2011).

[30] J. Leskovec and C. Faloutsos. "Sampling from large graphs". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006).

[31] J. Leskovec, J. Kleinberg, C. Faloutsos, H. D. Management, and D. Applications. "Graphs over time: densification laws , shrinking diameters and possible explanations". In: *Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, New York, USA: ACM Press, 2005.

[32] D. Li, Z. Xu, S. Li, and X. Sun. "Link prediction in social networks based on hypergraph". In: *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*. New York, New York, USA: ACM Press, 2013.

[33] H. Liang, Y. Xu, Y. Li, R. Nayak, and X. Tao. "Connecting users and items with weighted tags for personalized item recommendations". In: *Proceedings of the 21st ACM conference on Hypertext and hypermedia - HT '10*. New York, New York, USA: ACM Press, 2010.

[34] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. "Tag ranking". In: *Proceedings of the 18th international conference on World wide web - WWW '09* (2009).

[35] J. Liu, Z. Li, J. Tang, Y. Jiang, and H. Lu. "Personalized Geo-Specific Tag Recommendation for Photos on Social Websites". In: *IEEE Transactions on Multimedia* 16 (2014).

[36] M. Madden, A. Lenhart, S. Cortesi, U. Gasser, M. Duggan, A. Smith, and M. Beaton. *Teens, Social Media, and Privacy*. Tech. rep. Pew Research Center, 2013.

[37] A. Magnuson, V. Dialani, and D. Mallela. "Event Recommendation using Twitter Activity". In: *Proceedings of the 9th ACM Conference on Recommender Systems* (2015).

[38] C. S. Mesnage and M. J. Carman. "Tag navigation". In: *Proceedings of the 2nd international workshop on Social software engineering and applications - SoSEA '09*. New York, New York, USA: ACM Press, 2009.

[39] Nagehan Ilhan and S. G. Ogugucu. "A recommender model for social bookmarking sites". In: *Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control*. 2009.

[40] G. M. Namata, B. Staats, L. Getoor, and B. Shneiderman. "A dual-view approach to interactive network visualization". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management - CIKM '07*. New York, New York, USA: ACM Press, 2007.

[41] A. Nazir, A. Waagen, V. S. Vijayaraghavan, C.-N. Chuah, R. M. D'Souza, and B. Krishnamurthy. "Beyond friendship: Modeling User Activity Graphs on Social Network-Based Gifting Applications". In: *Proceedings of the 2012 ACM conference on Internet measurement conference - IMC '12*. New York, New York, USA: ACM Press, 2012.

[42] Ning Zhou, W. K. Cheung, Guoping Qiu, and Xiangyang Xue. "A Hybrid Probabilistic Model for Unified Collaborative and Content-Based Image Tagging". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33 (2011).

[43] O. Nov, M Naaman, and C. Ye. "What drives content tagging: the case of photos on Flickr". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2008).

[44] O Ozdikis, H Oguztuzun, and P Karagoz. "Evidential location estimation for events detected in twitter". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. 2013.

[45] F. Peregrino, D Tomás, and F Llopis. "Every move you make I'll be watching you: geographical focus detection on Twitter". In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*. 2013.

[46] T. Phan, J. Zhou, S. Chang, J. Hu, and J. Lee. "Collaborative Recommendation of Photo-Taking Geolocations". In: *Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia - GeoMM '14*. New York, New York, USA: ACM Press, 2014.

[47] P. Poseiro, M. Reis, C. Fortes, A. Sabino, and A. Rodrigues. "Aplicação do sistema HIDRALERTA de previsões e alerta de inundações: caso de estudo da Costa da Caparica". In: *Proceedings of the 3rd Jornadas de Engenharia Hidrográfica*. Lisbon, Portugal, 2014.

[48] P. Poseiro, A. Sabino, C. J. Fortes, M. T. Reis, and A. Rodrigues. "Aplicação do sistema HIDRALERTA de previsão e alerta de inundações: Caso de estudo da Praia da Vitória". In: *Proceedings of the 12th Congresso da Água*. 2014.

[49] M. Reis, P. Poseiro, C. Fortes, J. Conde, E. Didier, A. Sabino, and A. Rodrigues. "Risk Management in Maritime Structures". In: *Proceedings of the 8th International Conference on Management Science and Engineering Management*. Lisbon, Portugal, 2014.

[50] J. Rosa, A. Sabino, and A. Rodrigues. "Monitoring social network user density variations in areas of interest". In: *Proceedings of the 18th AGILE International Conference on Geographic Information Science*. 2015.

[51] M. Roth, A Ben-David, and D Deutscher. "Suggesting friends using the implicit social graph". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010).

[52] A Sabino and A Rodrigues. "A visual language for spatially aware agent-based modeling in crisis scenarios". In: *Proceedings of the 12th AGILE International Conference on Geographic Information Science*. 2009.

[53] A. Sabino and A. Rodrigues. "Productive Networks and Indirect Locations". In: *Citizen Empowered Mapping*. 2016.

[54] A. Sabino. "Agent-Based Simulation for Risk Management". Master Thesis. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2008.

[55] A. Sabino. "Space Aware Cooperative Environments". In: *Doctoral Consortium of the Collaboration Researchers International Working Group Conference*. 2010.

[56] A. Sabino and A. Rodrigues. "Understanding the role of cooperation in emergency plan construction". In: *Proceedings of the 8th International ISCRAM Conference*. 2011.

[57] A. Sabino and A. Rodrigues. "Indirect Location Recommendation". In: *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2014.

[58] A. Sabino, R. Nóbrega, A. Rodrigues, and N. Correia. "Life-Saver: Flood Emergency Simulator". In: *Proceedings of the 5th International ISCRAM Conference*. 2008.

[59] A. Sabino, A. Rodrigues, J. Gouveia, and M. Goulão. "Indirect Keyword Recommendation". In: *Proceedings of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence*. Warsaw, Poland, 2014.

[60] A. Sabino, J. Gouveia, and A. Rodrigues. "Visualizing Productive Network Relationships". In: *Proceedings of the 2014 IEEE/WIC/ACM International Conference on Web Intelligence*. Warsaw, Poland, 2014.

[61] A. Sabino, A. Rodrigues, J. Araújo, P. Poseiro, M. T. Reis, and C. J. Fortes. "Wave Overtopping Analysis and Early Warning Forecast System". In: *Proceedings of the 14th International Conference on Computational Science and Its Applications*. Ed. by B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi. Lecture Notes in Computer Science. Guimarães, Portugal: Springer International Publishing, 2014.

[62] A. Sabino, A. Rodrigues, P. Poseiro, M. T. Reis, C. J. Fortes, and R. Reis. "Coastal Risk Forecast System". In: *Proceddings of the 1st International Conference on Geographic Information Systems Theory, Applications and Management*. 2015.

[63] A. Sabino, J. Gouveia, and A. Rodrigues. "Visualizing productive networks". In: *IADIS International Journal on WWW/Internet* 12 (2015).

[64] A. Sabino, P. Poseiro, A. Rodrigues, M. T. Reis, C. J. Fortes, R. Reis, and J. Araújo. "Coastal Risk Forecast System". In: *Journal of Geographical Systems* pending pu (2017).

[65] M. Sachan, D. Contractor, T. Faruquie, and V. Subramaniam. "Probabilistic model for discovering topic based communities in social networks". In: *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*. New York, New York, USA: ACM Press, 2011.

[66] H. M. Sergieh, G. Gianini, M. Döller, H. Kosch, E. Egyed-Zsigmond, and J.-M. Pinon. "Geo-based automatic image annotation". In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval - ICMR '12*. New York, New York, USA: ACM Press, 2012.

[67] B. Sigurbjörnsson and R. Van Zwol. "Flickr tag recommendation based on collective knowledge". In: *Proceeding of the 17th international conference on World Wide Web - WWW '08* (2008).

[68] C. Smith. *By the numbers: 120+ Interesting Instagram Statistics*. 2015.

[69] M. Smith, D. L. Hansen, and E. Gleave. "Analyzing Enterprise Social Media Networks". In: *2009 International Conference on Computational Science and Engineering*. IEEE, 2009.

[70] K. Stefanidis, N. Shabib, K Nørvåg, and J. Krogstie. "Contextual recommendations for groups". In: *Advances in Conceptual Modeling: ER 2012 Workshops* (2012).

[71] V. Vijay and I. J. Jacob. "Combined approach of user specified tags and content-based image annotation". In: *2012 International Conference on Devices, Circuits and Systems (ICDCS)*. IEEE, 2012.

[72] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. "On the evolution of user interaction in Facebook". In: *Proceedings of the 2nd ACM workshop on Online social networks - WOSN '09*. New York, New York, USA: ACM Press, 2009.

[73] C. Wan, B. Kao, and D. W. Cheung. "Location-sensitive resources recommendation in social tagging systems". In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*. New York, New York, USA: ACM Press, 2012.

[74] Z. Wang, J. Feng, C. Zhang, and S. Yan. "Learning to rank tags". In: *Proceedings of the ACM International Conference on Image and Video Retrieval - CIVR '10* (2010).

[75] S. Wasserman and K. Faust. *Social network analysis methods and applications*. Cambridge University Press, 1994.

[76] L.-Y. Wei, M.-Y. Yeh, G. Lin, Y. H. Chan, and W. J. Lai. "Discovering Point-of-Interest Signatures Based on Group Features from Geo-social Networking Data". In: *2013 Conference on Technologies and Applications of Artificial Intelligence*. IEEE, 2013.

[77] Y.-T. Wen, K.-J. Cho, W.-C. Peng, J. Yeo, and S.-w. Hwang. "KSTR: Keyword-Aware Skyline Travel Route Recommendation". In: *2015 IEEE International Conference on Data Mining*. IEEE, 2015.

[78] C. Wilson, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. "Beyond Social Graphs: User Interactions in Online Social Networks and their Implications". In: *ACM Transactions on the Web* 6 (2012).

[79] C. Wu and B. Zhou. "Analysis of tag within online social networks". In: *Proceedinfs of the ACM 2009 international conference on Supporting group work - GROUP '09*. New York, New York, USA: ACM Press, 2009.

[80] G. Xu, Y. Zong, P. Jin, R. Pan, and Z. Wu. "KIPTC: a kernel information propagation tag clustering algorithm". In: *Journal of Intelligent Information Systems* 45 (2015).

[81] H. Xu, X. Zhou, M. Wang, Y. Xiang, and B. Shi. "Exploring Flickr's related tags for semantic annotation of web images". In: *Proceeding of the ACM International Conference on Image and Video Retrieval - CIVR '09* (2009).

[82] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. "On the semantic annotation of places in location-based social networks". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*. New York, New York, USA: ACM Press, 2011.

[83] W. Zhang, J. Wang, and W. Feng. "Combining latent factor model with location features for event-based group recommendation". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*. New York, New York, USA: ACM Press, 2013.

[84] T. C. Zhou, H. Ma, M. R. Lyu, and I. King. "UserRec: A User Recommendation Framework in Social Tagging Systems." In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

[85] C. Zhuang, Q. Ma, and M. Yoshikawa. "Location familiarity based flickr photographer classification for POI mining". In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*. New York, New York, USA: ACM Press, 2015.

[86] A. Zubiaga, C. Körner, and M. Strohmaier. "Tags vs shelves: from social tagging to social classification". In: *Proceedings of the 22nd ACM conference on Hypertext and hypermedia* (2011).

APPENDIX

PRODUCTIVE NETWORKS SURVEY RESULTS

Table A.1: Productive networks included in the survey. Network types are Content Indexing Network (CIN), Content Sharing Network (CSN), and Social Network (SN). The initial set of networks is signaled as seed.

Network	Type	Content type	Address	Seed
Academia.edu	CIN	Scientific articles	http://www.academia.edu	×
ACM Portal	CIN	Scientific articles	http://portal.acm.org	-
Allrecipes	CSN	Recipes	http://www.allrecipes.com	-
Arxiv	CIN	Scientific articles	http://arxiv.org	-
Asana	CSN	Tasks	http://www.asana.com	-
Blogger	CSN	Blog posts	http://blogger.com	-
Del.ici.ous	CSN	Urls	http://delicious.com	×
Endomondo	SN	Sport updates	http://www.endomondo.com	-
Facebook	SN	Social updates	http://www.facebook.com	×
Flickr	CSN	Photographs and images	http://www.flickr.com	×
Foursquare	SN	Place reviews	http://www.foursquare.com	×
Github	CSN	Source code	http://github.com	-
Goodreads	CSN	Book reviews	http://www.goodreads.com	×
Google Plus	SN	Social updates	http://plus.google.com	×
Hi5	SN	Social updates	http://www.hi5.com	×
IEEE Explore	CIN	Scientific articles	http://ieeexplore.ieee.org	-
IMDB	CSN	Movie ratings and reviews	http://www.imdb.com	-
Instagram	CSN	Photographs and images	http://instagram.com	×
ISI WoS	CIN	Scientific articles	http://webofknowledge.com	-
Last.fm	CSN	Music recommendation	http://htwww.last.fm	×
LinkedIn	SN	Professional experience	http://www.linkedin.com	×
LiveJournal	CSN	Blog posts	http://livejournal.com	×
Mendeley	CIN	Scientific articles	http://www.mendeley.com	-
Myspace	SN	Social updates	http://www.myspace.com	×
Picasa	CSN	Photographs and images	http://picasa.google.com	-
Pinterest	CSN	Personal interests	http://www.pinterest.com	×
ResearchGate	SN	Scientific research update	http://www.researchgate.net	-
Runkeeper	SN	Sport updates	http://runkeeper.com	-
Science Direct	CIN	Scientific articles	http://www.sciencedirect.com	-
Scridb	CSN	Books	http://www.scribd.com	-
Slack	CSN	Tasks and documents	http://slack.com	-
Slashdot	CIN	News	http://slashdot.org	-
Springer	CIN	Scientific articles	http://www.springer.com	-
Trello	CSN	Tasks	http://trello.com	-
Tumblr	CSN	Blog posts	http://www.tumblr.com	×
Twitter	SN	Social updates	http://twitter.com	×
Vimeo	CSN	Videos	http://vimeo.com	-
VK	SN	Social updates	http://vk.com	-
Wordpress	CSN	Blog posts	http://wordpress.com	-
Yelp.com	CIN	Business ratings	http://www.yelp.com	-
YouTube	CSN	Videos	http://www.youtube.com	-

Table A.2: Characterization of the users on the productive networks. When available, we present the estimated number of users declared by the network, and the Alexa global rank. Relationships between users may be direct of in the context of a group.

			Relationships	
Network	Number of users	Alexa global rank	Direct	Group
Academia.edu	18000000	806	×	×
ACM Portal	-	6494	-	-
Allrecipes	8000000	530	×	-
Arxiv	-	5589	-	-
Asana	-	527	×	×
Blogger	540000000	92	×	-
Del.ici.ous	9000000	1491	×	-
Endomondo	20000000	6793	×	×
Facebook	1280000000	2	×	×
Flickr	32000000	124	×	×
Foursquare	20000000	943	×	-
Github	3400000	96	×	-
Goodreads	13000000	275	×	×
Google Plus	1600000000	249254	×	×
Hi5	80000000	1962	×	×
IEEE Explore	-	788188	-	-
IMDB	59000000	43	-	-
Instagram	150000000	26	×	×
ISI WoS	-	6378	-	-
Last.fm	30000000	1362	-	×
LinkedIn	200000000	14	×	×
LiveJournal	36000000	175	×	×
Mendeley	2500000	17521	×	×
Myspace	50600000	1576	×	×
Picasa	-		-	-
Pinterest	70000000	32	×	×
ResearchGate	2300000	1237	×	×
Runkeeper	30000000	8687	×	×
Science Direct	-	903	-	-
Scridb	100000000	432	-	-
Slack	73000	704	-	×
Slashdot	5500000	1359	-	-
Springer	-	1418	-	-
Trello	5000000	362	×	-
Tumblr	50000000	31	×	-
Twitter	650000000	8	×	×
Vimeo	100000000	146	×	×
VK	230000000	36	×	×
Wordpress	409000000	33	×	-
Yelp.com	139000000	131	-	-
YouTube	1000000000	3	×	×

Table A.3: Primary media types of the productive networks. We distinguish URLs from general text.

	Media type			
Network	Text	Image	Video	URL
Academia.edu	×	-	-	-
ACM Portal	×	-	-	-
Allrecipes	×	×	-	-
Arxiv	×	-	-	-
Asana	×	-	-	-
Blogger	×	×	×	×
Del.ici.ous	-	-	-	×
Endomondo	×	-	-	-
Facebook	×	×	×	×
Flickr	×	×	-	-
Foursquare	×	-	-	-
Github	×	-	-	-
Goodreads	×	-	-	-
Google Plus	×	×	×	×
Hi5	×	×	×	×
IEEE Explore	×	-	-	-
IMDB	×	-	-	-
Instagram	×	×	-	-
ISI WoS	×	-	-	-
Last.fm	×	×	×	×
LinkedIn	×	-	-	×
LiveJournal	×	×	×	×
Mendeley	×	-	-	-
Myspace	×	×	×	×
Picasa	×	×	-	-
Pinterest	×	×	-	×
ResearchGate	×	-	-	-
Runkeeper	×	-	-	-
Science Direct	×	-	-	-
Scridb	×	-	-	-
Slack	×	-	-	-
Slashdot	×	-	-	×
Springer	×	-	-	-
Trello	×	-	-	-
Tumblr	×	×	×	×
Twitter	×	×	-	×
Vimeo	×	-	×	-
VK	×	×	×	×
Wordpress	×	×	×	×
Yelp.com	×	×	-	×
YouTube	×	-	×	-

Table A.4: Approach to keywords and annotation for each productive network.

Network	Keywords				
	Used for description	Used for classification	Separate from items	Available taxonomy	Users can create
Academia.edu	-	×	×	×	×
ACM Portal	-	×	×	×	×
Allrecipes	-	×	×	×	-
Arxiv	-	×	×	×	×
Asana	-	×	×	×	×
Blogger	×	-	×	-	×
Del.ici.ous	-	×	×	-	×
Endomondo	×	×	×	×	-
Facebook	×	×	-	-	×
Flickr	×	×	×	-	×
Foursquare	-	×	×	×	-
Github	-	×	×	×	×
Goodreads	-	×	×	×	-
Google Plus	×	×	-	-	×
Hi5	×	×	-	-	×
IEEE Explore	-	×	×	×	×
IMDB	×	×	×	×	×
Instagram	×	×	-	-	×
ISI WoS	-	×	×	×	×
Last.fm	×	×	×	×	×
LinkedIn	-	×	×	×	×
LiveJournal	×	×	-	-	×
Mendeley	-	×	×	×	×
Myspace	×	×	-	-	×
Picasa	×	×	×	-	×
Pinterest	×	×	-	-	×
ResearchGate	-	×	×	×	×
Runkeeper	×	×	×	-	-
Science Direct	-	×	×	×	×
Scridb	-	×	×	×	×
Slack	-	×	×	×	×
Slashdot	-	×	×	×	-
Springer	-	×	×	×	×
Trello	-	×	×	×	×
Tumblr	×	-	×	-	×
Twitter	×	×	-	-	×
Vimeo	×	-	×	-	×
VK	×	×	-	-	×
Wordpress	×	-	×	-	×
Yelp.com	-	×	×	×	-
YouTube	×	-	×	-	×

Table A.5: Approach to locations for each productive network. Networks may enable users to annotate items with place information, and even provide support for geographic coordinates.

Network	Locations				
	Enables places	Enables coordinates	Represents points	Represents areas	Separate from items
Academia.edu	×	-	-	-	×
ACM Portal	×	-	-	-	×
Allrecipes	×	-	-	-	×
Arxiv	×	-	-	-	×
Asana	-	-	-	-	-
Blogger	×	-	-	-	×
Del.ici.ous	-	-	-	-	-
Endomondo	×	×	×	×	×
Facebook	×	×	×	-	×
Flickr	×	×	×	-	×
Foursquare	×	×	×	×	×
Github	-	-	-	-	-
Goodreads	×	-	-	-	×
Google Plus	×	×	×	-	×
Hi5	×	-	-	-	×
IEEE Explore	×	-	-	-	×
IMDB	×	-	-	-	×
Instagram	×	×	×	-	×
ISI WoS	×	-	-	-	×
Last.fm	×	-	-	-	-
LinkedIn	×	-	-	-	×
LiveJournal	×	-	-	-	×
Mendeley	×	-	-	-	×
Myspace	×	-	-	-	×
Picasa	×	×	×	-	×
Pinterest	×	-	-	-	-
ResearchGate	×	-	-	-	×
Runkeeper	×	×	×	×	×
Science Direct	×	-	-	-	×
Scridb	-	-	-	-	-
Slack	-	-	-	-	-
Slashdot	×	-	-	-	-
Springer	×	-	-	-	×
Trello	-	-	-	-	-
Tumblr	×	-	-	-	×
Twitter	×	×	×	-	×
Vimeo	×	-	-	-	×
VK	×	×	×	-	×
Wordpress	×	-	-	-	×
Yelp.com	×	×	×	×	×
YouTube	×	-	-	-	×

Table A.6: Available search focus for each productive network.

	Search			
Network	Users	Items	Keywords	Locations
Academia.edu	×	×	×	-
ACM Portal	×	×	×	×
Allrecipes	×	×	×	×
Arxiv	×	×	×	-
Asana	×	×	×	-
Blogger	×	×	×	-
Del.ici.ous	×	-	×	-
Endomondo	×	×	×	-
Facebook	×	×	×	×
Flickr	×	×	×	-
Foursquare	-	×	×	×
Github	×	×	×	-
Goodreads	×	×	×	-
Google Plus	×	×	×	-
Hi5	×	×	×	×
IEEE Explore	×	×	×	-
IMDB	×	×	×	-
Instagram	×	×	×	-
ISI WoS	×	×	×	-
Last.fm	-	×	×	×
LinkedIn	×	×	×	×
LiveJournal	-	×	×	×
Mendeley	×	×	×	-
Myspace	×	×	×	×
Picasa	×	×	×	-
Pinterest	×	×	×	-
ResearchGate	×	×	×	-
Runkeeper	×	×	×	-
Science Direct	×	×	×	-
Scridb	-	×	×	-
Slack	×	×	×	-
Slashdot	×	×	×	-
Springer	×	×	×	-
Trello	×	×	×	-
Tumblr	-	×	×	-
Twitter	×	×	×	×
Vimeo	×	×	×	-
VK	×	×	×	×
Wordpress	×	×	×	-
Yelp.com	-	×	×	×
YouTube	×	×	×	-

Table A.7: Author and user relationships.

	Author Relationships			User Relationships		
Network	Direct	Bi-direct	Group	Direct	Bi-direct	Group
Academia.edu	×	×	×	×	×	×
ACM Portal	×	×	×	-	-	-
Allrecipes	-	-	-	×	-	-
Arxiv	×	×	×	-	-	-
Asana	×	×	×	×	×	×
Blogger	×	×	-	×	-	-
Del.ici.ous	-	-	-	×	-	-
Endomondo	-	-	-	×	×	×
Facebook	×	×	×	×	×	×
Flickr	-	-	-	×	-	×
Foursquare	-	-	-	×	×	-
Github	×	×	×	×	-	-
Goodreads	×	×	-	×	×	×
Google Plus	×	×	×	×	-	×
Hi5	×	×	×	×	×	×
IEEE Explore	×	×	×	-	-	-
IMDB	×	×	-	-	-	-
Instagram	-	-	-	×	-	×
ISI WoS	×	×	×	-	-	-
Last.fm	×	×	×	-	-	×
LinkedIn	-	-	×	×	×	×
LiveJournal	×	×	-	×	-	×
Mendeley	×	×	×	×	×	×
Myspace	×	×	×	×	-	×
Picasa	-	-	-	-	-	-
Pinterest	-	-	-	×	-	×
ResearchGate	×	×	×	×	×	×
Runkeeper	-	-	-	×	×	×
Science Direct	×	×	×	-	-	-
Scridb	×	×	×	-	-	-
Slack	×	×	×	-	×	×
Slashdot	-	-	-	-	-	-
Springer	×	×	×	-	-	-
Trello	×	×	×	×	×	-
Tumblr	×	×	-	×	-	-
Twitter	×	-	-	×	-	×
Vimeo	×	×	-	×	-	×
VK	×	×	×	×	×	×
Wordpress	×	×	-	×	-	-
Yelp.com	-	-	-	-	-	-
YouTube	×	-	-	×	-	×

Appendix B

Interaction and Visualization Platform Development and Evaluation

B.1 User Characterization

The interaction and visualization platform was evaluated by 15 users, 7 aged between 21 and 25 years old, 1 aged between 31 and 40 years old, 4 aged between 41 and 55 years old, and 1 over 55 years old. All users had college level education, or were engaged in undergraduate studies. 11 users work or study in information technology areas.

B.2 First Questionnaire - First Part: User Trial Script

Users were given up to 5 minutes to interact with the platform, without any introduction on specific operators or concepts, after which the evaluation would initiate. Each user trial was guided by the following script:

1. *Please identify a photograph, and what you see that is directly related with it.*
2. *Please chose a photograph, identify its keywords, and chose one or more you find interesting. Please explain your understanding of the procedure.*
3. *Select a keyword. By doing that, you are able to visualize all links with other keywords. Please explain the difference in thickness of the lines. Select one link. Please explain what happened next. May you add one photograph of that link to the visualization?*
4. *Please select a photograph and execute a keyword expansion. Please explain what happened. Visualize the content of a group and remove a photograph into the visualization.*
5. *Please find a photograph with a keyword. Find a photograph related with it.*

6. *Please find photographs that are relevant to an user.*

7. *Now imagine that you are the owner of the photographs in the visualization.*

    a) *How would you find photographs of other users which share interests with you?*

    b) *How would you find users which share interests with you?*

    c) *How would you find photographs associated with a particular keyword?*

Figure B.1 presents the time required to answer each question.

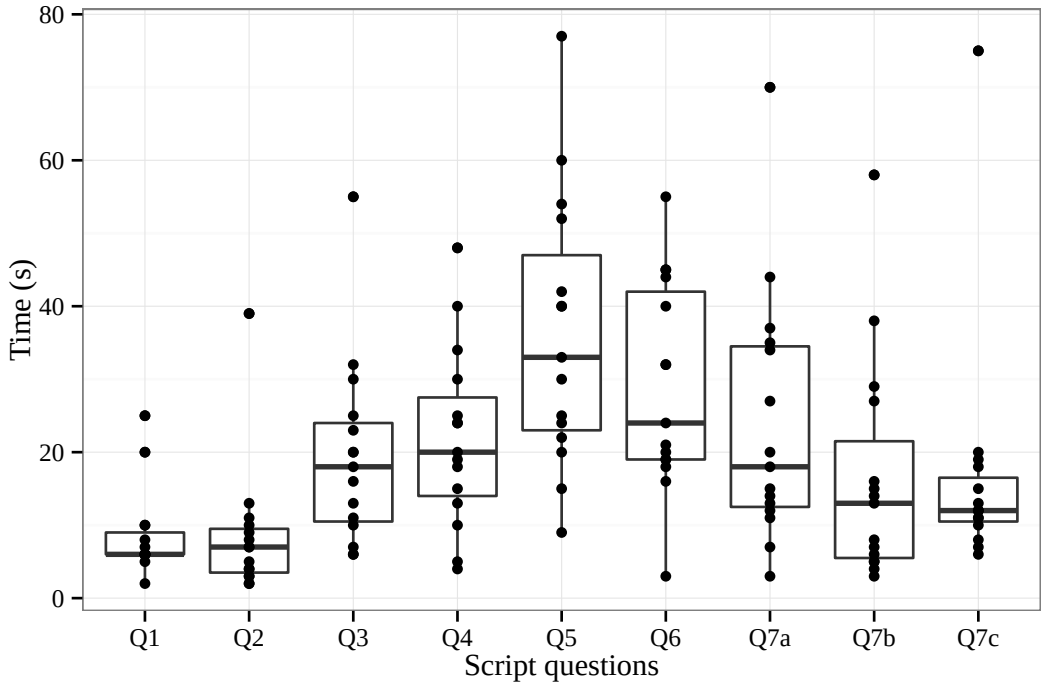

Figure B.1: User answer times for each question of the script presented in B.2.

B.3 First Questionnaire - Second Part: General Assessment Answer Scales

The first questionnaire included a part designed to assess user general appreciation of the platform. The questionnaire was delivered in Portuguese. Table B.1 presents the translation and the original text.

Table B.1: The general assessment questionnaire. The original Portuguese text and rating semantic differential scale terms is presented in parenthesis with each question.

Q1. How would you describe the application? (*Como melhor descreveria a aplicação?*)

Intuitive (*Intuitiva*)	1 2 3 4 5	Complicated (*Complicada*)
Unpleasant (*Desagradável*)	1 2 3 4 5	Pleasant (*Agradável*)
Practical (*Prática*)	1 2 3 4 5	Unpractical (*Incómoda*)
Creative (*Criativa*)	1 2 3 4 5	Ordinary (*Banal*)
Structured (*Bem estruturada*)	1 2 3 4 5	Confusing (*Confusa*)
Boring (*Aborrecida*)	1 2 3 4 5	Engaging (*Cativante*)
Ugly (*Feia*)	1 2 3 4 5	Pretty (*Bonita*)
Useless (*Inútil*)	1 2 3 4 5	Useful (*Útil*)
Evident (*Clara*)	1 2 3 4 5	Cryptical (*Críptica*)

Q2. Please rate the need of an expansion button. (*Classifique a necessidade do botão de expansão no programa.*)

Useless (*Nada útil*) 1 2 3 4 5 Very useful (*Muito útil*)

Q3. How useful if the grouping of nodes for cluttering reduction? (*Quão útil é o agrupamento de nós para a redução de excesso de informação?*)

Useless (*Nada útil*) 1 2 3 4 5 Very useful (*Muito útil*)

Q4. Do you consider the visualization change operation useful? (*Considera que a operação de mudança de visualização é útil?*)

Useless (*Nada útil*) 1 2 3 4 5 Very useful (*Muito útil*)

Q5. Please rate how easy it was to identify the current context. (*Classifique a facilidade que teve em identificar em que contexto se encontrava?*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q6. Please rate how easy it was to navigate to the desired context. (*Classifique a facilidade que teve em navegar até ao contexto que desejava.*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q7. Please rate how easy it was to understand the meaning of a node in a visualization. (*Classifique a facilidade em definir o significado de um nó numa das visualizações.*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q8. Please rate how easy it was to understand the meaning of an edge in a visualization. (*Classifique a facilidade em definir o significado de uma ligação numa das visualizações.*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q9. Please rate how easy if was to identify the strongest connections (edges). (*Classifique a facilidade de identificação das ligações mais fortes*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q10. Please rate how easy it was to use the visualization change operation. (*Classifique a facilidade de utilização da operação de mudança de visualização*)

Hard (*Difícil*) 1 2 3 4 5 Easy(*Fácil*)

Q11. Was it dificult to navigate in the program, or to find the desired information? (*Sentiu alguma dificuldade na navegação do programa, ou em encontrar informação necessária?*)

Yes (*Sim*) No (*Não*)

Q11. a) If you answered yes in the previous questions, please explain the problem and in which operations. (*Se respondeu sim à pergunta anterior, por favor explique qual foi a dificuldade sentida e em que operações a sentiu.*)

Free text answer.

Q12. Are there any unsupported connections which you would like to identify through the system? (*Existem ligações que gostaria de identificar através do sistema e que não são possíveis de identificar?*)

Yes (*Sim*) No (*Não*)

Q12. b) If you answered yes in the previous questions, please elaborate. (*Se respondeu sim à pergunta anterior, por favor explique.*)

Free text answer.

Appendix C

Set-Builder Notation

A set is an unordered list of elements, which is defined by logical predicates. In this documents, set-building expressions are presented using the following notation

$$\{x \mid \Phi(x)\}$$

where x is a variable, the symbol "|" is a vertical separator which reads "*such that*", and $\Phi(x)$ is a logical predicate. All values of x where the predicate evaluates to *true* belong to the set.

Logic predicates may include several terms, in which case the terms are separated by the symbol "$\wedge$", which reads "*and*".

Most set-building expressions begin with a term that indicates the domain of the variable,

$$\{x \mid x \in X \wedge \ldots\}$$

which is usually followed by the remaining terms of the logical rule.

C.1 Complex Predicates

When required, terms may include other variables, subject to specific conditions, which constraint the x variable. These are represented by

$$\{x \mid x \in X \wedge \exists y \in Y : (\Phi(x, y))\}$$

where y is a second variable, the symbol ":" represents a second vertical separator, reading "*such that*", and $\Phi(x, y)$ is a logical predicate applied to both variables. Predicates in these conditions are enclosed by parenthesis to avoid confusion.

C.2 Writing Long Predicates

When possible, predicates with several terms are represented in several lines, grouped with curly braces. The expression

$$\{x \mid x \in X \land \Phi_1(x) \land \Phi_2(x)\}$$

is equivalent to

$$\{x \mid x \in X \land \left\{ \begin{array}{l} \Phi_1(x) \\ \Phi_2(x) \end{array} \right\}\}$$

C.3 Selecting Elements From Sets

Expressions to select elements from sets may include inequality constraints represent as subscripts. The expression

$$X_i, X_j \in X : X_i \neq X_j$$

is equivalent to

$$\underset{i \neq j}{X_i, X_j \in X}$$

Appendix D

Systematic Literature Review Protocol

Our literature review, presented in chapter 2, is guided by the following protocol:

1. Define which publication indexing databases to query;
2. Identify query expressions;
3. Collect results from all databases, for each query expression;
4. Rule out results based on publication title;
5. Rule out results based on abstract;
6. Rule out results based on the full document content.

Queries are explicitly executed to search for publication keywords. To rule out publications based on their title, we evaluate if they refer to different research domains, but which may use the same terminology as ours, such as:

- Computer networks architecture
- Psychology of groups
- Graphs outside social networks

To rule out publications based on the abstract we identify whether it describes research unrelated with our research questions. Finally, to rule out publications based on document content we take into account the following criteria:

- The publication is ultimately unrelated with our research questions;
- It is a preliminary result of another publication already in the collection;

- Presents a weak argument, e.g., lacking a detailed description of the proposal, or presenting non significant evaluation.

We considered three main publication indexing services:

ACM digital library (ACMDL) A major collection of full-text publications and bibliographical records covering many research topics in the field of computer science. Available at `http://dl.acm.org/`

SCOPUS database of peer-reviewed literature (SCOPUS) Presently the largest bibliographical reference database of peer-reviewed publications. Available at `https://www.scopus.com/`

IEEE Xplore digital libraty (IEEEX) Another important collection of full-text publications and bibliographical records in the field of computer science, electrical engineering and electronics. Available at `http://ieeexplore.ieee.org/Xplore/home.jsp`

Each query was queried with a combination of at least two keywords. The following expressions returned positive results on at least one database:

QE 1 Tagging, User Density Estimation

QE 2 Social Graph, User Density Estimation

QE 3 Tagging, Location Recommendation

QE 4 Social Graph, Tagging, Location Recommendation

QE 5 Insight, Social Networks, Social Graph

QE 6 Perception, Social Networks, Social Graph

Table D.1 presents a quantitative description of our review collection.

The final set of articles includes publications between 2005 and 2016, with 60% dating from 2012 or after. This collection was further extended with publications we became aware of in several others contexts, such as conference attendance or peer recommendation. See chapter 2 for the complete review.

Table D.1: Number of publications in the collection, from the initial query results to the final set.

Query	Database	Result	Title Check	Abstract Check	Final
QE 1	ACMLDL	4	2	1	1
	IEEEX	9	5	3	3
	SCOPUS	2	1	1	1
QE 2	ACMLDL	1	1	1	1
	SCOPUS	12	4	3	3
	IEEEX	2	0	0	0
QE 3	ACMLDL	61	47	16	14
	SCOPUS	3	2	2	2
	IEEEX	17	9	7	6
QE 4	ACMLDL	1	1	0	0
	SCOPUS	5	1	0	0
	IEEEX	4	3	1	1
QE 5	ACMLDL	105	37	15	13
	SCOPUS	30	20	7	7
	IEEEX	3	0	0	0
QE 6	ACMLDL	12	2	0	0
	SCOPUS	5	2	0	0
	IEEEX	18	1	1	1
Total		294	138	58	53