UNIVERSITY OF LJUBLJANA
FACULTY OF COMPUTER AND INFORMATION SCIENCE

Anita Valmarska

ANALYSIS OF CITATION NETWORKS

DIPLOMA THESIS

UNIVERSITY STUDY PROGRAMME

COMPUTER AND INFORMATION SCIENCE

Mentor: Assoc. Prof. Janez Demšar, PhD

Ljubljana, 2014

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Anita Valmarska

ANALIZA OMREŽIJ CITATOV

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

MENTOR: izr. prof. dr. Janez Demšar

Ljubljana, 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 01947 / 2013
Datum: 5.9.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **ANITA VALMARSKA**

Naslov: **ANALIZA OMREŽIJ CITATOV**
ANALYSIS OF CITATION NETWORKS

Vrsta naloge: DIPLOMSKO DELO UNIVERZITETNEGA ŠTUDIJA

Tematika naloge:

Analiza omrežij je eno najbolj živahnih področij podatkovnega rudarjenja, ki se uporablja na najrazličnejših področjih znanosti, od biologije do analize socialnih omrežij. V diplomski nalogi raziščite možnosti uporabe mrež citiranj v bibliometriki. Konkretno, zanima nas, ali je mogoče z analizo citatov določiti vpliv posameznih člankov na razvoj področja prek različnih pristopov oziroma paradigem. V diplomi izberite primerno področje (npr. psihologijo), zberite informacije o člankih in sestavite mrežo citatov ter mrežo na primeren način skrčite in analizirajte.

Mentor: Dekan:

izr. prof. dr. Janez Demšar

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

prof. dr. Nikolaj Zimic

Izjava o avtorstvu diplomskega dela

Spodaj podpisana Anita Valmarska, z vpisno številko **63060404**, sem avtor diplomskega dela z naslovom:

Analysis of citation networks

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelala samostojno pod mentorstvom izr. prof. dr. Janeza Demšarja,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 18. marca 2014 Podpis avtorice:

Zahvaljujem se svojemu mentorju, izr. prof. dr. Janezu Demšarju, za ideje, čas ter potrpežlivost pri izdelavi diplomske naloge. Želim se zahvaliti dr. Matiji Svetini za njegove koristne nasvete.

Iskrena hvala družini za vso potporo in nenehno spodbudo.

I would like to thank my mentor, dr. Janez Demšar, for his ideas, his time, and his patience during the course of writing this thesis. I would also like to thank dr. Matija Svetina for his valuable advice.

My sincerest appreciation goes to my family, for their tireless support and continuous encouragement.

Contents

Povzetek

V diplomski nalogi raziskujemo problem odkrivanja raziskovalnih disciplin v izbrani znanosti. Odkrivanje raziskovalnih disciplin je osnovano na raziskavi povezav med izbrano množico člankov in članki, ki jih citirajo. V naši diplomski nalogi smo se odločili obravnavati članke s področja psihologije. Psihologija je v pogledu tega dela zanimiva znanstvena disciplina, ker obravnava večje število različnih tem. Ameriško psihološko združenje (*American Psychological Association*, APA) je razdelilo področje raziskav iz psihologije na 54 različnih poddisciplin. Pri pisanju diplomske naloge nas je zanimalo, ali lahko rekonstruiramo to razdelitev na podlai objavljenih člankov in njihovih medsebojnih citatov.

Prvi korak pri izdelavi diplomske naloge je bilo zbiranje podatkov. Ker iskanje centralizirane baze podatkov, ki bi ustrezala našim pogojem, ni bilo uspešno, smo se odločili, da bomo osnovno množico člankov na področju psihologije dobili preko avtomatskega brskanja straneh Wikipedije. Wikipedija je prosta enciklopedija na internetu, katere članki so organizirani po kategorijah in temah, ki jih obdelujejo. Začeli smo s kategorijo Psihologija [1] na angleški Wikipediji in preiskali vse kategorije in članke do globine 5. Ker se uredniki Wikipedije ne držijo enotnega standarda za citiranje, smo morali izbrati enoten način za zbiranje znanstvenih člankov. Odločili smo se obdržati le članke, pri katerih je bil podan DOI. S tem smo seveda izgubili precej člankov, ki so morda imeli velik vpliv na razvoj Psihologije, vendar je bil tak pristop neizogiben, če smo hoteli priti do zanesljivega avtomatskega zbiranja podatkov. Pri iskanju v globino je skripta našla tudi znanstvene članke, ki niso bili povezani s psihologijo, temveč so pripadali kategorijam,

[1] http://en.wikipedia.org/wiki/Category:Psychology

ki so z njo le posredno povezane, kot na primer religija, biologija, sociologija itd.

Po začetnem zbiranju osnovnih člankov iz psihologije smo se usmerili k zbiranju podatkov o znanstvenih člankih, njihovih avtorjih, revijah, kjer so bili objavljeni ter podatke o vseh člankih, ki so osnovne članke navedli v njihovem razdelku z literaturo. Na začetku smo imeli težave pri izbiranju iskalnika, ki bi ponujal informacije o citiranju člankov. Na koncu smo se odločili za Microsoftov iskalnik MAS[2] (Microsoft Academic Search). To je prosti iskalnik, ki ponuja iskanje po DOI, je prost za uporabo in ima lepo organizirane profile avtorjev, člankov in revij. Za vsakega od člankov, ki smo jih dobili z iskanjem po Wikipediji, smo poiskali podatke, ki jih opisujejo, in podatke o člankih, ki jih citirajo.

Dobljene podatke smo predstavili v obliki usmerjene mreže. Tako smo dobili mrežo citatov. Mreže citatov so aciklične usmerjene informacijske mreže, kjer mrežna struktura odraža strukturo informacij, ki so shranjeni v vozliščih mreže. Vozlišča mreže predstavljajo znanstveni članki in usmerjena povezava iz članka A do članka B pomeni,da članek A citira članek B. Začetna analiza dobljene mreže je razkrila lastnosti, ki so značilne tudi za mreže iz resničnega. Ena med njimi je, na primer, porazdelitev stopenj v mreži, ki v mreži citatov, podobno kot pri drugih mrežah, sledi potenčnem zakonu.

Problem odkrivanja raziskovalnih disciplin iz znanstvenih citatov na področju psihologije smo prevedli na problem gručenja mreže, oziroma na problem odkrivanja skupin v mreži. V našem primeru smo ga reševali z uporabo algoritma Louvain. To je algoritem za odkrivanje skupin v mreži, ki je hiter, enostaven in učinkovit. Louvain deli vozlišča mreže v različne skupine in pri tem zahteva optimizacijo modularnosti. Prvotna uporaba algoritma je lepo ločila članke iz psihologije od ostalih znanstvenih člankov. Dobili smo veliko povezano komponento, ki je vsebovala zbrane raziskave na področju psihologije, ter veliko število manjših skupin, ki niso bile povezane med seboj. Podrobnejši pregled je pokazal, da gre za skupine člankov s področija, kot so biologija, religija, geologija itd.

Ponovna uporaba algoritma Louvain nad veliko komponento je vrnila 52 skupin. Zaradi velikosti in zahtevnosti problema ni bilo mogoče ročno, samo s pregledom

[2]http://academic.research.microsoft.com/

vozlišč in njihovih naslovov, oceniti uspešnosti razdelitve. Zato smo se odločili, da bomo primerjali naše rezultate z razdelitvijo, ki jo je podala APA, z opazovanjem kosinusne podobnosti med najdenimi članki in referenčnimi besedili iz vsakega področja APA.

Preden smo se lahko lotili s primerjalnim postopkom, smo morali najprej zgraditi bazo referenčnih besedil za vsako področje APA. Za vsako področje APA na svoji strani navede po eno revijo, v kateri se tipično objavljajo članki, ki sodijo na to področje. Te revije smo poiskali na strani MAS in zbrali vse članke, ki so bili objavljeni v njih. Za revije, ki niso imele svojega zapisa na MAS, smo poiskali revije, ki so jim bili zelo podobni. Iz naslovov člankov iz vsake revije smo zgradili dokumente z besedami, ki so dajali sliko o pogostih besedah vsakega področja. Izločili smo besede, kot so predlogi in vezniki, ter uporabili tf-idf (frekvenca terminov, inverzna frekvenca dokumenta) s ciljem, da bi pravilno odrazili vpliv vsake besede.

S primerjavo med naslovi člankov in dokumenti, ki so predstavljali izraze, ki so pogosti na posameznih področjih psihologije, smo za vsak članek dobili imena najbolj podobnih področij ter seznam tem, ki jih ta področja raziskujejo. Na podlagi dobljenih imen in raziskovalnih tem smo po lastni presoji določili imena skupin. Prikazali smo poimenovane skupine in povezave med njimi.

Rezultati poimenovanja in povezave med določenimi skupinami kažejo na relativno dobro rekonstrukcijo razdelitve področja psihologije. Pri ocenjevanju teh rezultatov moramo biti seveda previdni: zavedati se moramo, da smo pri poimenovanju skupin uporabili subjektivno presojo, kar ima velik vpliv pri končni oceni kakovosti porazdelitve mreže.

Rezultati dajejo spodbudo za nadaljnje raziskove in izboljšave. Možne so izboljšave pri vsakem koraku izdelave diplomske naloge. V prihodnosti se lahko odločimo, da se pri zbiranju člankov omejimo zgolj na najvplivnejše revije. Pri gradnji baze člankov in njihovih citatov lahko gledamo obojestransko in zbiramo še o podatkih o člankih, ki so jih naši članki citirali. Možne so tudi izboljšave pri postopku poimenovanja.

Ključne besede: teorija mrež, mreže citatov, gručenje grafov, odkrivanje skupnosti, kosinusna podobnost.

Abstract

In this thesis we explore the problem of detection of research subdisciplines of a chosen science, based only on the data about citing papers from a chosen batch of papers relevant to the selected science. We directed our attention to the field of Psychology. It is an interesting scientific discipline, with variety of research topics and numerous scientific publications throughout the years.

Due to lack of freely accessible centralized database of psychological papers and their relevant citations, the first step of our thesis was collection of papers and their applicable citations. Data was presented in form of a citation network. Citation networks are acyclic directed information networks, where the structure of the network reflects the structure of the information stored in the network vertices. The process of differentiation of research disciplines in the network was performed by applying a state-of-the-art algorithm for community detection. Our choice was the Louvain method, known for its simplicity, effectiveness and speed.

Part of the mechanism for rating the detected communities was to name them and examine their connections. Due to the vast quantity of available data and the unfamiliarity with the psychological field, we named the communities based on the measures for cosine similarity between our initially collected psychological papers and the relevant texts for each of the APA divisions of Psychology.

Results obtained by the network analysis and the method for community detection are positive. However, the nature of data collection and the influence of our subjective judgment for community naming offer lots of opportunities for further improvement.

Keywords: network theory, citation networks, graph clustering, community detection, cosine similarity.

Chapter 1

Introduction

Networks are a part of our everyday life. Entities of any type that share some kind of a relationship form networks. Examples include the connected network of airports from around the world, a simple family network or protein-protein interactions on cellular level. In real-world networks, the interaction between entities of the network does not happen by accident. The interaction formation follows certain patterns/norms significant for the type of entities and the type of network. For example, it is more likely to form a friendship with a person if that person is already a friend of some of your friends. The network structure can offer us a deep understanding of the network itself, the roles of the entities, the dynamics of the network, and a sound platform for predicting how even small changes can affect the network.

Citation networks are directed networks in which one paper cites another. Reasons for citations are many. In most cases the authors cite older publications in order to identify the related body of work, to substantiate claims or establish precedence, or to legitimate own statements or assumptions. In the scientific world citations are used also to critically analyze or correct earlier work. Since the scientific output of authors is often quantitatively measured by the number of received citations, it is not unfamiliar for authors to cite their older work or the work of their collaborators.

Intuitively, it can be expected for papers to cite more often other papers that

are published in the same research discipline, and within the same science. It is orderly for papers to enlist in their bibliography section publications that have already been established as claims in the researched discipline. We were interested to find out whether we could detect research disciplines from a single science based only on the its citation network. For the purpose of this thesis we chose Psychology.

Psychology is an academic and applied discipline that involves the scientific study of mental functions and behaviors. It is a discipline that traces its philosophical roots back to the ancient civilizations of Egypt, Greece, China, India, and Persia. Psychology studies border on the line with studies on various other fields including physiology, neuroscience, sociology, anthropology, as well as philosophy and other components of the humanities. It is a discipline that constantly develops and explores numerous aspects of the human mental functions. Psychologists divide their attention among various sub-fields of psychology. The researched topics in Psychology are numerous, and the intensity with which they are explored many times depends on the current condition of the society. Throughout the years the topic of marriage and divorce has been an interesting and always enigmatic topic for psychologists to explore. Family is one of the pillars of society, and understanding the psychological manner of its functioning can help the whole society to move forward.

Our goal was to explore whether we can identify disciplines and topics of psychology only from a citation network of papers published in the field of psychology. The intention was to apply one of the state-of-the-art algorithms for community detection in the hope that we would be able to differentiate among the disciplines and topics of psychology.

In chapter 2 we offer a brief overview of network theory. In chapter 3 we discuss the problems and limitations of collecting data about psychological papers. Next we overview the nature of citation networks. In chapter 5 we discuss the problems of network clustering and we introduce the Louvain method. This is followed by presentation of our results and a conclusion in chapter 7.

Chapter 2

Network theory

Network theory is applied science of discrete mathematics and it is part of graph theory. Beginnings of graph theory reach to the 18th century when Euler gave the solution of the seven bridges of Königsberg problem (*Figure 2.1*). It is a science that has greatly developed, both theoretically and practically. Networks are indispensable when we are interested in the relationships between entities. They are very useful for pattern discovery in relationships we are interested in. Examples of network usage can be found in sociology, chemistry, biology, physics, computer science, economics and many other areas. Examples include the Internet, the Wide World Web, social networks of acquaintances or other form of individual interactions, organizational networks, business networks of relationships between companies, neutral networks, network of blood vessels as one of the representatives of distribution networks and many more (*Figure 2.2*).

Many aspects of the mentioned network systems are worthy of study. Some people study the nature of individual entities e.g. how a company functions or how a human being acts or feel in the system she is placed in. Others research the nature of connections and interconnections in the systems, for example, the protocols that are used on the Internet or the dynamics of business relationships between companies. Recent years have witnessed new movement in network research, where the focus have shifted away from the analysis of small graphs and the properties of individual vertices and edges within such graphs to consideration

of large-scale statistical properties of graphs. This new movement explores the pattern of connection between the entities of the networks and its influence on the behavior of the studied system. The new approach has been driven mostly by the availability of computers and information networks that allow us to gather and analyze data on a scale much larger than previously possible. Networks of tens or hundreds vertices have been replaced by networks with million or even billion vertices. The size of available networks have forced change to the analytical approach. Many of the questions previously asked no longer have meaning and their answers do not give a sufficient explanation for the behavior of the researched system [24].

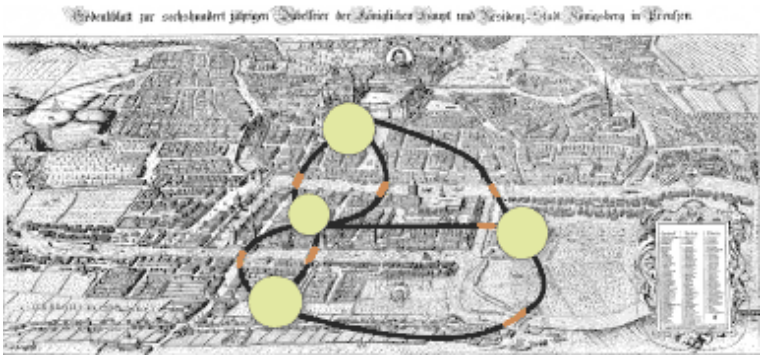

Figure 2.1: *A popular pastime among the Königsberg's residents was to look for a path through town that traversed all of its seven bridges without crossing the same bridge twice. Euler recognized that the solution had nothing to do with any of the dices involved, but rather with the way in which the landmasses were connected to each other. He abstracted the problem by assigning each destination a letter and denoted the bridges by pairs of letters connecting two destinations. He proved that there is no possible path by which one can visit every destination, crossing every bridge only once. His proof from 1735 is today known as finding an Euler path in a graph and it is the first real proof in graph theory.*

Network N is defined by two sets, V and E, $N := (V, E)$. V is a set of nodes, and E is a set of edges that represent the relationship between the nodes from V.

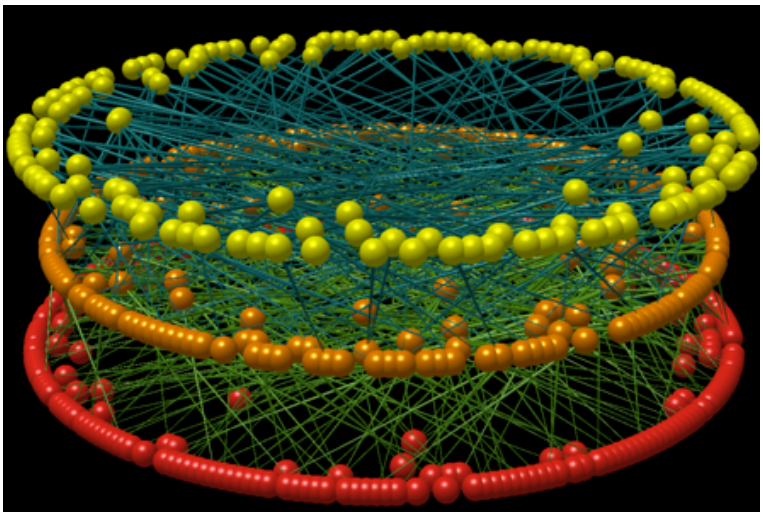

Figure 2.2: *Network analysis of hyperdiverse and specialized parasitoid food webs. This is an Apanteles wasp food web. There are 143 species of wasps (top ring), 203 species of host caterpillars (middle ring) and 266 species of caterpillar food plants (bottom ring). Janzen/Hallwachs data from ACG identifies 419 unique tri-trophic interactions. The illustration shows that most microgastrine parasitoid wasps only attack 1-2 species of host caterpillar(s) and most caterpillars only eat a few species of host plants. Picture courtesy of Neo Martinez and Richard Williams* [16].

The edges in E may be directed or undirected.

$$V = \{v_1, v_2, v_3, ..., v_n\} \tag{2.1}$$

$$E \subseteq \{\{v_i, v_j\} | v_i, v_j \in V\} \tag{2.2}$$

Equations 2.1 and *2.2* are the same as the definition of graphs in discrete mathematics. In network theory, we can define a network by assigning properties to its nodes and edges. Both the nodes and the edges between them can be of different types and have different weights. Therefore, in a citation network where the nodes represent publications, each node can be assigned unique properties, e.g. the year of publication, the number of authors, the publisher etc.

Initial definition of networks is extended by introduction of directed edges. In the field of citation networks, the edge direction is very important. Ordered pair

of nodes (v_i, v_j) represents a direct-oriented relationship from node v_i to node v_j.

$$E \subseteq \{(v_i, v_j) | v_i, v_j \in V\} \tag{2.3}$$

Scientists in wide variety of fields have developed an extensive set of tools, mathematical, computational and statistical, for analyzing, modeling and understanding of networks. Many of these tools concentrate on small networks and by simple calculations provide information that may be interesting for the user, for example, pointing to the vertex with most connections, shortest path between two vertices etc. Other tools concentrate on providing mathematical methods for predicting the processes and the behavior of networks. The mathematical form of these tools tries explain the Internet traffic flow or how a disease will spread through the community. These are abstract models that can be used for networks that represent different systems. However, a well-posed question and basic understanding of the researched field are necessary in order to obtain a clearer picture of the system dynamics.

2.1 Network properties

2.1.1 Small world effect

Probably the best known phenomenon in network theory is the *small world effect*. The phenomenon was initially observed by the social scientist Milgram in 1967 in his attempt for quantifying the typical distance between entities in social networks. The "geodesic distance" between two vertices in a network is the minimum number of edges that must be traversed from one vertex to the other through the network. Mathematical arguments suggest that distance should be quite short between most of the vertices in most of the networks. He wanted experimentally to test this conjecture on real networks.

Milgram sent a set of packages, 96 in all, to recipients randomly chosen from the telephone directory in the US town of Omaha, Nebraska. He gave his observed entities the task to form relations and transfer the small packages from Omaha

to Boston, without using the postal service. Results were surprising. Most of the packages arrived at their destinations and on average less than six interchanges were needed. This research has showed that society is a network that has the property of small world. Every pair of vertices in that network is connected by relative short distance. Following the results from this experiments, scientists adopted the term *six degrees of separation* which is even today widely used in the network theory literature [21, 24].

During this experiment a number of packages were lost on their way. This is the argument that number of scientists have used to oppose the experiment and the obtained results.

We measure the small world phenomenon by observing the average distance between two vertices. Let n be the number of vertices in the researched network and l the average shortest path between every pair of vertices in the network. We calculate the value of l using the equation

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum d_{ij} \tag{2.4}$$

where d_{ij} represents the shortest distance between vertices v_i and v_j. If the vertices are not connected then $d_{ij} = 0$. *Equation 2.4* gives us a good measure of whether the phenomenon is present in a chosen network. Results from scientific researches in recent years have confirmed that this phenomenon is present in networks with several thousand or even million nodes.

The small-world effect affects the dynamics of processes taking place on networks. For example, if we consider the spread of information or across a network, the small-world effect implies that the spread will be fast in most of the real-world networks.

2.1.2 Degree distribution

Network researchers have dedicated a great attention to the degree distribution in real-world networks. Node degree represents the number of edges incident to a

chosen node. In directed graphs we distinct between *in-degree distribution*, number of edges directed towards a node, and *out-degree distribution*, number of edges pointing out from a node. Numerous researchers have shown that degree distribution in real-world networks differs from the degree distribution in random networks and in regular networks.

Let p_k be the probability that the degree of a randomly chosen node in a random network equals k. The degree distribution in random network where each of the $\frac{1}{2}n(n-1)$ possible edges is present with probability p is binomial

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2.5}$$

This differs significantly from the real-world networks. Numerous studies have shown that degree distribution is skewed to the right. This indicates that real-world networks include vertices with degrees that are much higher than the average vertex degree. However, the total number of such vertices is small, and it cannot be used for statistical evaluation of the distribution for very big k.

Real-world networks have degree distributions that follow the power law, *Equation 2.6*, with $2 < \alpha < 3$. Networks whose degree distribution follows the power law are called *scale-free networks*.

$$p_k \sim k^{-\alpha} \tag{2.6}$$

2.1.3 Transitivity

A clear deviation from the behavior of the random graph can be seen in the property of network transitivity. In many networks it is found that if a vertex A is connected to a vertex B, and the same vertex B is connected to a vertex C, than it is highly probable that the vertex A is also connected to the vertex C. In the terms of social networks this can be understood as that the friend of your friend is likely to be your friend. The transitivity ratio can be quantified by the equation:

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of connected triples of vertices}}, \tag{2.7}$$

where a "connected triple" denotes a single vertex with edges running to an unordered pair of others. The transitivity ratio C measures the fraction of triangles among all pairs of connected triples. The factor of three is used because each triangle contributes to the formation of three triangles.

Watts and Strogatz proposed an alternative definition of the transitivity ratio which is also widely used [37]. Their proposition includes definition of a local value

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i}. \tag{2.8}$$

The transitivity ratio of the whole network of size n is then

$$C = \frac{1}{n} \sum_i C_i \tag{2.9}$$

2.2 Network models

One of the best ways to understand how network structure and its properties influence the behavior of the network is to build mathematical models. A good way to address this question is to construct artificial networks that mimic the properties of the researched network and observe the relationship between the properties and the dynamics of the network. This is the rationale behind the construction of random networks. They shed light on the structural properties of networks and are widely used to model dynamical processes on networks. Bellow we list two types of network models

2.2.1 Poisson random networks

This is a very simple network model independently proposed by Solomonoff and Rapoport [33], and the famous mathematician Paul Erdős and Alfred Rényi [7, 8].

The random graph is the perfect example of a good mathematical definition: it is simple, has surprisingly intricate structure, and yields many applications.

The Erdős-Rényi model $G_{n,p}$ is defined as a graph with n vertices, where each pair of vertices is connected with probability p. Technically $G_{n,p}$ is the ensemble of all such graphs in which a graph with m edges appears with probability $p^m(1-p)^{M-m}$, where $M = \frac{1}{2}n(n-1)$ is the maximum possible number of edges. When $n \to \infty$ the degree distribution is Poisson, defined as

$$p_k \approx \frac{z^k e^{-z}}{k!}, \tag{2.10}$$

where $z = p(n-1)$.

The random graph reproduces well one of the principle features of real-world networks, namely the small-world effect. The mean number of neighbors at a distance l from a vertex in a random graph is z^d, and hence the value of d needed to encompass the entire network is $z^l \simeq n$. Thus a typical distance through the network is $l = \frac{\log n}{\log z}$, which satisfies the definition of the small world. However, in almost all other respects, the properties of the random graph do not match those of the real world. It has a low transitivity ratio. The probability of connecting a pair of vertices is p, regardless of whether they have a common neighbor. The system also has a Poisson degree distribution, unlike the distributions of real-world networks. It has random mixing patterns, it does not show any correlation between degrees of adjacent vertices, has no community structure and the navigation through this systems is impossible using local algorithms [21].

2.2.2 The small world model

The small-world model is a less sophisticated model with high transitivity, proposed by Watts and Strogatz [37]. Networks are systems that may have a geographical component to them. Vertices of the network have positions in space and it can be a reasonable assumption that the geographical proximity may have a role in deciding which pairs of nodes are connected. The small-world model starts from this idea by positioning a network built on a low-dimensional regular lattice and

then adding or moving edges with the intention of creating shortcuts that connect parts of the lattice that are otherwise remote to one another.

The best studied case so far is the small-world model built on one-dimensional lattice of L vertices with periodic boundary conditions, a ring. We join each vertex to its neighbors k or fewer lattice spacings away and we get a system with Lk edges. The small-world model is then created by moving the existing edges to other positions in the lattice. The rewiring process allows creation of a small-world networks, with properties interpolated between a regular lattice and a random graph. When $p = 0$, we have a regular lattice where the transitivity ratio is $C = (3k - 3)/(4k - 2)$, $C = 3/4$ for large k. The mean geodesic distances between vertices tend to $L/4k$ for large L. When $p = 1$, every edge is rewired to a new, random location and the resulting graph is almost a random graph with geodesic distances tending to $\log L/\log k$, and a very small transitivity ratio $C \simeq 2k/L$. Transformation of a regular lattice to an almost random network through the process of rewiring is shown on *Figure 2.3*.

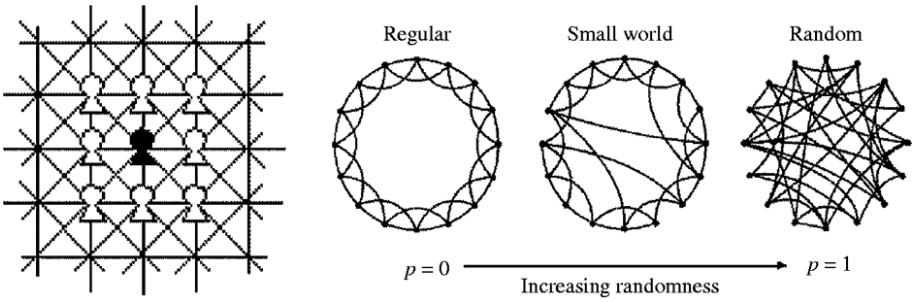

Figure 2.3: *Network transformation from a regular lattice to an almost random graph. Small-world networks interpolate the properties from the two network models.*

2.3 Real-world networks

Interest in the study of networks has significantly increased in the last 20 years. The groundbreaking research was made by the physicists. Their research perspective differs from the one adopted by mathematicians or sociologists. Physicists conduct empirical research on real-world networks - Internet, friendship networks, biological systems etc. In contrast to the research proceeded by sociologists, physicists are interested in the statistical properties of networks. This new way of thinking has resulted in discovery of very interesting network properties which were not possible in the previous type of research. This has lead to development of new theories, algorithms, models and metrics which describe the dynamics of network systems.

The continuous development of network theory is due to the vast knowledge collected over the years from the scientific research on the fields of mathematics, computer science, social sciences, biology, bio-chemistry etc.

Real-world networks can be roughly divided into four sets [21].

Biological networks are mostly networks from the fields of molecular biology, genetics and neurology. An interesting example of biological network, which is attracting a lot of interest in the scientific world is the food web. It is a network where the vertices represent species in an ecosystem and the directed edge between species A and the species B represents that species A preys on species B. In terms of energy flow, the edges of the network can be reversed, so that the relationship between the predator and the prey is represented by directed edge from the prey B to predator A.

Technological networks are man-made networks designed for distribution of some commodity or resource, such as electricity or information. Examples of these networks are the electric power grid, telephone and delivery networks, roads network, railway network etc. Special example of the technological networks is the Internet. The ever changing number of computers connected to the Internet and the structure of their connection represents an exciting challenge for the scientists. For practical reasons, the analysis of the Internet is limited to the analysis of the network of interconnected routers.

In the technological networks the geographical location of its entities has a very important role. Entities that are closer locally have a bigger probability of forming a relationship.

Social networks describe a set of people and their relationships and interactions. Part of this set is any network that describes friendships between individuals, partner relationships between individuals and companies, paper co-authorships etc. Information networks and social networks are the sets of real world networks that have been researched the most in the past. Traditional social networks studies often suffer from problem of inaccuracy, subjectivity and small sample size. Data collection is a problem that causes headaches for the researchers of social networks. It is usually done by querying participants directly using interviews and questionnaires. It is a labor-intensive work that often limits the size of the observed network. Many times, data obtained by the surveys is affected by subjective biases on the part of the respondents; ones definition and understanding of the term "friendship" may differ from another.
The data collection problems have turned researchers to other methods of probing social networks. One source of reliable data are the collaboration networks. These are networks in which participants collaborate in groups of one kind or another. Typical and widely known example of collaboration network is the collaboration network of actors which is documented on the Internet Movie Database[1].
Social sciences such as sociology, psychology and economics offer the basics for development of social networks analysis.

Information networks: typical representative of this set of real-life networks are the citation networks of scientific papers. The vertices represent different scientific papers and are connected by directed edges. Two vertices are connected if one paper cites another. Citation networks are acyclic since papers can only cite scientific papers that have already been written and published, and not papers that are about to be written, thus preventing the formation of cycles. Directed edges in citation networks point backwards in time, offering knowledge about the papers that have started a certain scientific trend. Network structure represents

[1] www.imdb.com/

information that is encoded into the vertices.

An important representative of the information networks is also the network of the World Wide Web. Web-pages are represented by vertices of the network, and there is a directed edge from vertex A to vertex B only if the web-page that is represented by the vertex A contains a hyperlink that points to the web-page represented by vertex B.

Chapter 3

Data collection

3.1 Data collection problems

The first step into constructing a network of any kind is data collection. Being novice into the field of psychology we had to complete a small research about the available citation databases for Psychology.

As far as our knowledge reaches, there is no central database containing publications in the field of Psychology. Consequently, we decided for scrawling the pages connected with Psychology on Wikipedia. Wikipedia is a collaboratively edited, multilingual, free Internet encyclopedia that is supported by the non-profit Wikimedia Foundation. It is a web service that is well known to the average web user, and it is constantly expanding, offering more extensive and more precise information about a wide variety of topics. Wikipedia has become instrumental for the introducing and attraction of new followers to certain trends, themes etc. The average user usually tries to find information about concepts that are new to her on Wikipedia and consequently decides whether to expand her knowledge with information from third parties. Frequently these third party sources are listed in the reference section of Wikipedia pages.

Wikipedia content is organized in the form of hierarchy. Important topics are gathered into categories. Each category is divided into subcategories and pages which are, according to Wikipedia, directly connected to the upper categories.

Each page on Wikipedia offers information about the topics it describes. These pages are known as Wikipedia articles. In order to support the claims made on Wikipedia, authors of the articles must cite reliable third party sources. They must offer inline citations for any material that is challenged or likely to be challenged, and for any quotation in any part of the Wikipedia articles. Any unsourced material runs the risk of being challenged or eventually removed from the encyclopedia. Wikipedia prompts its editors to use consistent citation format on each page. At their last resource, they should offer at least enough information for successful identification of the source.

References are listed in the dedicated section of each page. The relative citation freedom granted by Wikipedia allows its editors to be uncoordinated in regard to the citation format used in articles on Wikipedia. Different articles use different formats and it is not uncommon for different citation formats to be used in the same article. In many cases even non-standard formats are used. It is nearly impossible to construct a good program that will be able to successfully extract information only about the relevant academic publications used as third party source for verification of the written content.

Once we decided to use Wikipedia as our initial source into the field of Psychology, we proceeded with collecting the information about psychology publications from the reference section of the articles connected to the Psychology category[1] on Wikipedia in English. Due to the citation inconsistencies mentioned above, we decided to extract only the reference material identified by DOI (*Digital Object Identifier*).

The DOI system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers for use on digital networks. Metadata and URL where the electronic document can be found in association with its DOI. The DOI system has been developed and implemented in a range of publishing applications since 2000. DOI names can identify creative works such as texts, images, audio or video items, and software in both electronic and physical forms, performances, and abstract works such as licenses, parties to a transaction,

[1] http://en.wikipedia.org/wiki/Category:Psychology

etc. The names can identify different level of detail. A DOI name can identify a journal, an individual issue of a journal, individual article or even a single table in an article. There are numerous registration agencies which are responsible for issuing DOIs and updating the metadata associated with them. The multilingual European DOI registration agency, mEDRA, and the Chinese registration agency, Wanfang Data, are responsible for the publications in non-English language markets.

Our choice to use DOI for publication identification meant that many of the older and most influential publications on the field of Psychology will be omitted from the initial data selection. By later inspection, we discovered that the initial DOI extraction did not include the academic works of some of the most prominent contributors in the development of Psychology, like Jean Piaget, Sigmund Freud, B. F. Skinner, Kurt Koffka, Max Wertheimer, Wolfgang Köhler and others.

Decision to adopt the DOI system for initial acquisition of Psychology papers was based on the assumption that articles that use the DOI for publication identification carry more importance and are created by authors that have some substantial scientific background, in our case a deeper knowledge in Psychology. We believed that this compromise would result in extraction of the most prominent publications in each of the fields of Psychology and extraction of some of the publications that have shaped the development of psychology through the years, from the teachings of the Ancient Greek philosophers, through Skinner's theory about behaviorism to today's modern ideas and conclusions in Psychology. Unfortunately, as stated above, this was not the case. However, our research discovered some interesting patterns and conclusions which are presented in 'Data analysis'.

3.2 Citation tools

Once we collected the set of DOIs connected with the category Psychology on Wikipedia, we needed to find a suitable citation tool. A suitable citation tool in our case is a citation tool that includes academic publications from the field of Psychology, one that is free, allows usage of a crawling script and allows for

searching by DOIs. We tried several tools and most of them did not satisfy one or more of the conditions mentioned above. The Science Citation Index[2] (along with its sister publications, the Social Science Citation Index[3] and the Arts and Humanities Citation Index[4]) do not offer citation search by publication DOI. It is the same case with CiteSeerX[5] and the Social Science Research Network[6].

PsycINFO[7] is a database of abstracts of literature in the field of psychology. It contains citations and summaries from the 19th century to the present of journal articles, book chapters, books, and dissertations. It is database that is weekly updated and by October 2013 it contained over 3.5 million records. However, it is a tool that requires a subscription and therefore could not be used in our work.

Google Scholar[8] (GSC) is the "academic" version of the popular Google search engine. It covers academic literature from different sources, including "academic publishers, professional societies, online repositories, universities and other web sites" (Google, 2011). It is the largest academic search engine that besides the scientific papers available online, also harvests other academic materials, court opinions and patents.

Each profile on GSC should be personally self-created and self-edited by the author though Google personal account. The author can select her own references, her partners and the labels that best describe her fields of interest in a free natural language. It is an unrestricted model that grants the users the complete ownership over their profiles and allocates them the capabilities to freely edit and modify them [26].

Google does not disclose the names of these sources or the frequency of updates. This means that we cannot be certain how comprehensive a search is or how up-to-date it is. GSC is a free, fast, simple and easy to use tool. It allows access to a wider and larger audience, GSC allows publication search by DOI and it enlists a

[2]http://ip-science.thomsonreuters.com/
[3]http://thomsonreuters.com/social-sciences-citation-index/
[4]http://thomsonreuters.com/arts-humanities-citation-index/
[5]http://citeseerx.ist.psu.edu/
[6]http://www.ssrn.com/en/
[7]http://www.apa.org/pubs/databases/psycinfo/index.aspx
[8]http://scholar.google.com/

full list of publication's citations. However, it does not allow usage of a crawling script. Our script for data collection managed to extract information only for a handful of publications before it was identified as a crawler and blocked. Google's policy to block crawling scripts forced us to focus on Microsoft Academic Search[9] (MAS).

3.2.1 Microsoft Academic Search

Like GSC, MAS indexes millions of scholarly papers. It is an experimental research dataset developed by Microsoft Research to explore how scholars, scientists, students, and practitioners find academic content, researchers, institutions, and activities. Microsoft Academic dataset indexes not only millions of academic publications, it also displays the key relationships between and among subjects, content, and authors, highlighting the critical links that help define scientific research [41].

MAS is a free citation tool. It offers multidisciplinary databases of academic publications. MAS is a scientific web database which gathers bibliographic information from the principal scientific publishers (Elsevier, Springer) and bibliographic services (CrossRef). As of August 2012, it contains 40 million of documents. Profiles on MAS are automatically created from the names of the authors of these papers. Besides other units like journals, institutions or conferences, MAS also built profiles for individuals which includes the author's list of publications, bibliometric indicators (publications, citations), disciplinary areas of interest and different sets of the most frequent co-authors, journals and keywords. Each of the profiles on MAS includes an identification number which is randomly assigned [26].

MAS goes beyond document retrieval service that count citations. It automatically provides bibliographic records about authors, journals, institutions or research disciplines. Although it has limited quality control, it can still be considered valid for research evaluation and scientific benchmarking. Given the limitations in the control of identifiers, the most interesting feature is that the whole search service relies on self-edited personal profiles. That means that they can be

[9] http://academic.research.microsoft.com/

updated, modified or merged, after approval, by the researchers themselves. This relative freedom allows the scientists to boast a public, qualitatively controlled and accessible curriculum to spread their research activities and overall performance. However, as stated before, any additional modification requires a prior approval and unethical behavior is penalized [26].

MSA is updated weekly. It includes over 250 million citations, and can be used to visualize connections between documents, authors, conferences and journals. Users can make keyword searches or searches through publication DOI, author names, publication titles, journal titles, conference titles, organization names or their respective identification numbers.

Just a couple of papers have been published about the performance of MAS. Jacsó presented a review of its principal functionalities in comparison with Scopus and Web of Science, concluding that MAS may become a free tool to help the research assessment.

3.3 Data collection process

First objective in our data collection was to collect the DOIs related to the category Psychology on Wikipedia. As we mentioned in the previous section, categories on Wikipedia are organized in hierarchy, followed by subcategories and articles. Deeper in the hierarchical tree can be found articles that act as connectors between categories that at first can be considered as unconnected. A visit to the articles in the lower categories of the Psychology tree on Wikipedia reveals articles that are connected to Psychology, but can be considered to have a closer connection to other disciplines. Examples include pages connected to pages from Religion, Evolution, Biology, etc.

After some consideration we decided to go to the depth level of 5. The decision was based solely on the difference between the number of visited categories and articles on depth 4, depth 5 and depth 6. This included all subcategories and pages on Wikipedia which could be reached in 5 or less steps from the initial category Psychology. Subcategories and pages that are accessed by direct hyperlinks from

the category Psychology are on the level 0. As it was previously decided, from each retrieved subcategory and page on level depth to 5 we extracted the DOIs of corresponding cited publications.

Once this step was completed, we proceeded by querying the MAS for each of the collected DOIs. If a publication was found on MAS, we collected information about the title of the publication, its authors, the year of publication, the journal, ID of the publication, IDs of its authors, etc. Afterward, we proceeded with collecting the same information about the publications that have cited the queried publication. From the gathered information we built a database whose structure is shown in *Table 3.1.*

The additional information about each of the queried publications and the corresponding citing publications were collected with the intention and hope of future improvement of the current research. We would like to explore whether there are some patterns in the co-author network, how does the publication of a paper in a certain journal affects the paper's relevance, importance and the number of received citations. At the moment, we are intrigued to explore the formation of co-authorship relation on the basis of the simple geographic placement of the authors of the papers. Information about the year of publication can offer us an insight in the change of number of citations and the other trends of citation through time. At the moment, the simple information about where a certain paper has been published can assist us to narrow down the scientific research discipline it deals with.

A short draft of the data collection process can be found in *Figure 3.1.* As it is to be expected, this was a time consuming process. It involved understanding of the citation practices on Wikipedia, finding the most convenient citation tool, a close comprehension of the profile structure on MSA, accepting and respecting the MSA Terms of use and finally extraction of the desired information.

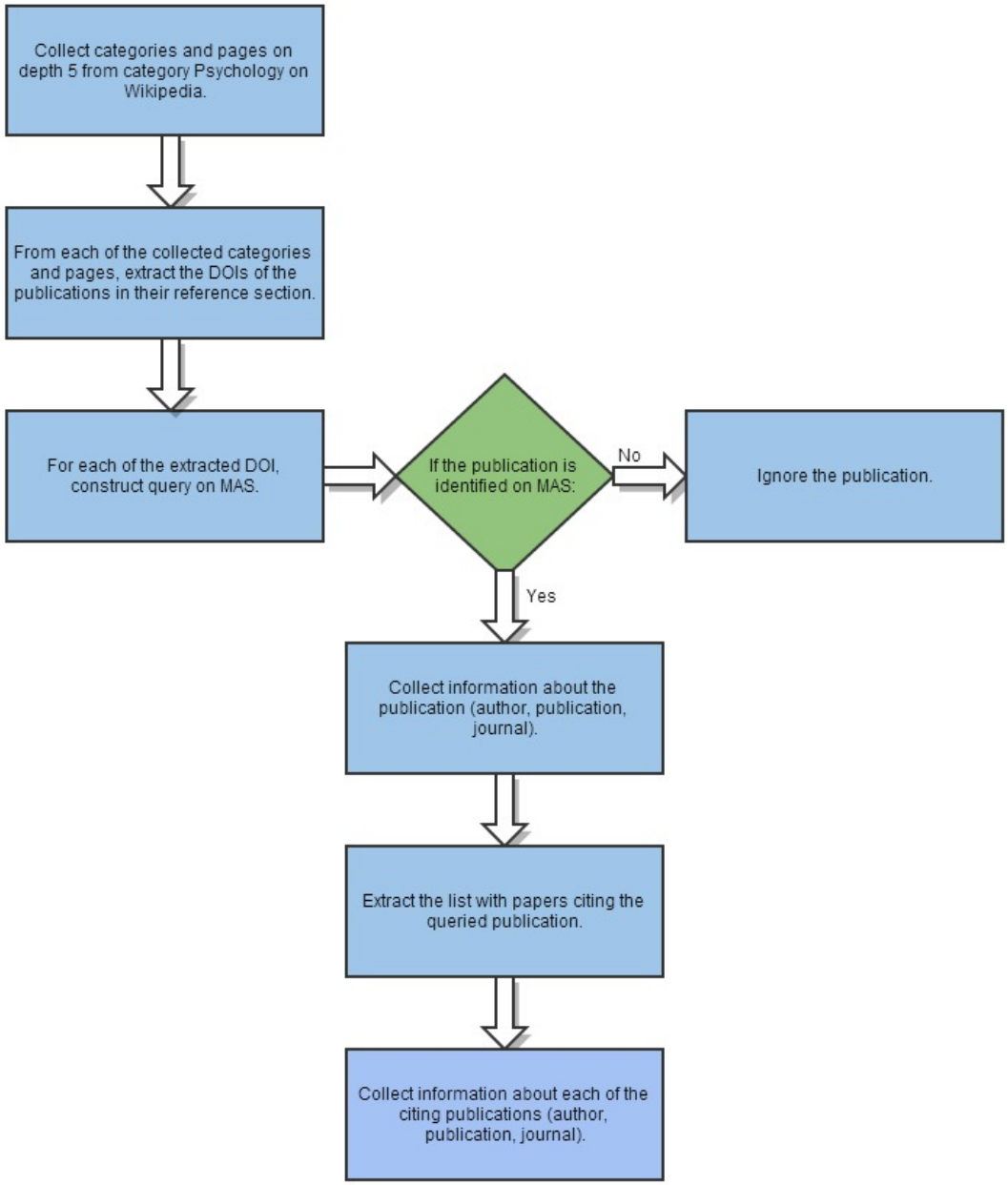

Figure 3.1: *Data collection process*

3.4 Statistics of the collected data

Wikipedia search reached a total of 3,173 visited categories. These categories had links to additional 71,606 unique articles on Wikipedia. DOI extraction from

Table	**Attributes**
`Authors`	Author's MAS ID, Author's name, Author's href on MAS
Papers	Paper's MAS ID, Year of publication, Journal MAS ID, Paper title, Paper's href on MAS, Citations href, No. of citations
`Journals`	Journal's MAS ID, Journal's title, Journal's href on MAS
`DOI to ID`	Paper's DOI, Paper's MAS ID

Table 3.1: *Database structure*

Wikipedia pages resulted in 63,826 unique DOIs. It corresponds to the number of queries created on MAS. As we explained above, for each of the extracted DOIs from Wikipedia we composed a query on MAS in search for the corresponding publications. For each of the found publications we saved the relevant information and the information about their citing publications. This process resulted in the total of 953,428 publications, written by 1,589,144 authors, and published in 12,862 journals. We constructed a citation network from the retrieved data. This network consisted of 948,791 vertices and 1,539,563 edges. The difference between the total number of publications present in the database and the number of papers included in the citation network represent the number of queried DOIs which at the time of retrieval had not received a single citation.

The oldest publication was published in 1773, while the newest were published in the year of 2013. Around 13% of the extracted publications were not identified by the year of publication. The average number of authors per paper was 4.92. This number is informative but does not reflect the actual result. It is MAS policy to list only the first 13 authors of a publication. Many of the publications in the field of medicine are written by a bigger number of authors, which involve collaboration of scientists from different fields of medicine, students, assistant teachers, etc. This limitation on the part of MAS influences the total number of authors in our database, number which is surely higher than the one presented here. Short statistics of the collected data is shown in *Table 3.2*.

No. of visited categories	3,173
No. of visited pages	71,606
No. of publications	953,428
No. of authors	1,589,144
No. of journals	12,862
Oldest publication (year)	1773
Newest publication (year)	2013
Publications without publication year (%)	13
Avg. no. of authors per publication (year)	4.92
Cit. network, no. of vertices	948,791
Cit. network, no. of edges	1,539,563

Table 3.2: *Short statistics of the collected data*

Chapter 4

Citation networks

Citation networks offer a special network representation of academic literature entities. Most papers refer to one or more previous papers, usually in the bibliography section at the end of the paper. This provides the capacity for one to construct a network with directed edges from one vertex to another. Vertices in a citation network represent some sort of documents, while directed edge from v_i to v_j denotes that document v_i has listed document v_j among the documents in its reference literature. This is done with the intention to direct the user to information that may be useful to her, to give credit for prior work, to indicate influences on current work or to disagree with the content of the cited academic paper. Citation of papers is a clear indication that contents of the earlier papers are in some kind of relationship with the content of the later one. Hence citation networks are networks of relatedness on subject matter.

Quantitative studies of citation networks reach back to the 1960s. Earliest known study of citation networks seems to be the study conducted by Price [27] in 1965. Studies of this kind fall within the field formerly known as "library science", but now it is more often referred as "information science". The branch of information science dealing specifically with the statistical study of publications and citations is called *bibliometrics*.

The most common way to assemble citation data is to do it manually, simply typing in all the entries from the bibliography section of papers to create a database

that later can be used used to construct the network. The earliest databases of this kind were the basis for development of what is now known as the Science Citation Index. The Science Citation Index is one of the primary and most common used sources of citation data. Its records are hand-maintained by professional staff and they offer reasonably complete and accurate coverage of literature. Citation data from the Science Citation Index is quite expensive to acquire.

In recent years citation indexing by computer has become more common. Most popular example is the website Citeseer, maintained by Pennsylvania State University. Citeseer performs citation indexing of papers in the field of computer science and information science by crawling the Web to find freely available manuscripts of papers in electronic form, and then searching through those papers to identify citations leading to other papers. The biggest detriment of computer indexing is the fact that many papers are not available online or they do not allow free access. Paper citations can be listed in different formats and may also include errors. One paper may exist in more than one place on the Web as well as in journals and books, and possibly in more than one different version.

Citation networks are in many ways similar to the World Wide Web. Vertices hold information in the form of text and pictures, just like web-pages, and the links from one paper to another play a role similar to hyperlinks, offering the reader a path to information that is relevant to the discussed content. Papers with many citations are often more influential and widely read than those with few, just as it is the case with web-pages. Citation networks can be "surfed" by following succession of citation from paper to paper, the same way computer users surf the Web [24].

However, there are numerous differences between the citation networks and the other real-world networks. Their edges are directed from one vertex to another, forming an acyclic relationship between vertices in the network. Mathematically we can represent citation network as a adjacency matrix, where each of its elements are defined as

$$A_{ij} = \begin{cases} 1 & \text{if there is a connection from } i \text{ to } j; \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

In directed networks the adjacency matrix is usually asymmetric.

Real citation networks are not always acyclic. There are cases when a scientific publication can cite future work. For example, an author can cite her other work which is in the process of being published. This can lead to cycles in the network. Usually these cycles are rare and they are limited to a certain time period.

Another important characteristics of the citation networks is their ability to change through time. New documents are added and the structure of citation networks evolves through time. The time development of citation networks takes a special form. Vertices and the edges that connect them are added at certain point in time, and can not be deleted. This means that the structure of citation networks is mostly static, and it changes only when a new document is added in the set of vertices of the network. Contrary to other information networks, for example, the World Wide Web, in citation networks we cannot delete nor modify the edges that were already added in the network. Their limited time development offers cleaner data for research of the network's time growth than the networks of the World Wide Web.

Citation networks reveal some interesting statistics. Around 47% of all papers in the Science Citation Index have never been cited at all. Of the remainder, 9% have one citation, 6% have two citations and only 21% percent of all papers have 10 or more citations. Just 1% of all the recorded papers in the Science Citation Index have 100 or more citations [24]. These figures can be explained by the power-law distribution discussed in chapter 2.

The most highly cited paper in the Science Citation Index is a methodological paper in molecular biology by Lawry *et. al* which has been cited over 250.000 times.

Citation networks are present in different fields. Most of the scientific literature dedicated to analysis of citation networks is focused on research of citations between scientific documents. However, part of the scientists are interested also

in research of the citation networks of product patents. Patents cite other patents for many reasons. Usually it is with the intention of establishing their uniqueness and originality that separates them from other, older patents. Patent data allow construction of enormous citation networks. In recent years scientists show great interest in the research of citation networks of legal documents. Relationships in these citation networks represent citations made by judges and other legal subjects on legal matters. Most of the time citation networks of legal documents are used for discovery of legal precedents.

Work in the field of citation networks have been compelling for numerous scientists. The first person who set the basic model and started to research citation networks was Price [27] in 1965. In his article *"Networks of scientific papers"*, published in *Science*, he tried to give answers for the layout of the world network of citations of scientific publications. Price focused his research on the citation occurrence, change over time of number of citations for a certain publication and the inability to predict the probability whether the existing number of citations can influence the increase of number of citations in the future. Price talks about the "immediacy factor" which is responsible for the well known phenomenon when scientific articles become outdated after 10 years.

In research of citation networks the biggest disagreement was by the conclusions in regards to the degree distribution. Currently we can isolate four groups of scientists asserting different degree distributions.

The first group of scientists [27, 29, 32], claim that the degree distribution in citation networks follows the power law. They insist that p_k, the probability that the degree of a randomly chosen vertex in a citation network is equal to k, can be determined by the *Equation 2.6*.

The second group of authors [30, 28] persist that degree distribution is described by the equation

$$C(k) = Ae^{-b\ln k - c(\ln k)^2} \tag{4.2}$$

where A, b and c present previously calculated constants.

Wallace *et al* [36] and Anastasiadis *et al* [1] have conducted several experiments on citation networks and they claim that the results from those experiments support their hypothesis that the degree distribution in citation networks follows the Tsallis' distribution. According to them, the number of publications that have received c citations can be calculated by the equation

$$N(C) \propto \frac{1}{[1+(q-1)\frac{c}{T}]^{\frac{1}{q-1}}}, q \simeq \frac{4}{3} \tag{4.3}$$

for all available citations c. T is the so-called "effective temperature".

Van Raan [34] in his work from 2001 claims that the degree distribution follows the modified Bassel's function.

A very intriguing task in the field of citation analysis is discovery of the most influential publications in a selected scientific field. *PageRank* is an algorithm for link analysis, introduced by Larry Page [3], that has had a considerable influence on the development of the web search engine Google. In 2007 Chen [4] suggested usage of the *PageRank* algorithm for classification of scientific papers in APS journals. The same year Walker [35] proposed modification of the algorithm, where it also considers the age of the publication when performing a classification of scientific publications.

Chapter 5

Community detection

Any nonuniform data contains underlying structure due to the heterogeneity of the data. The process of identifying this structure in terms of grouping the data elements is called clustering [15]. Community detection is a type of cluster finding problem in networks. It must not be confused with the problem of network partition [24]. Community detection problems differ from graph partitioning in that the number and the size of the groups into which the network is divided are not specified beforehand. The goal of community detection is to find the natural fault lines along which a network separates. The sizes of the groups may vary.

Community detection is a tool for analysis and understanding of network data. Knowledge of the group structure might help us understand the organization of the underlying system. *Figure 5.1* shows a very famous example in community detection literature. It displays a network of friendships between high-school students [18]. As it is evident from the figure, the presented friendship splits into two clear groups which are dictated by students' ethnicity. Structure of the researched network can shed a light onto the nature of the social interactions within the community presented.

Community detection has uses in other types of networks as well. For example, clusters of nodes in a web-graph can represent groups of related web-pages. Clusters of nodes in citation networks might indicate scientific research disciplines. Similarly, clusters of nodes in a metabolic network can reveal functional units

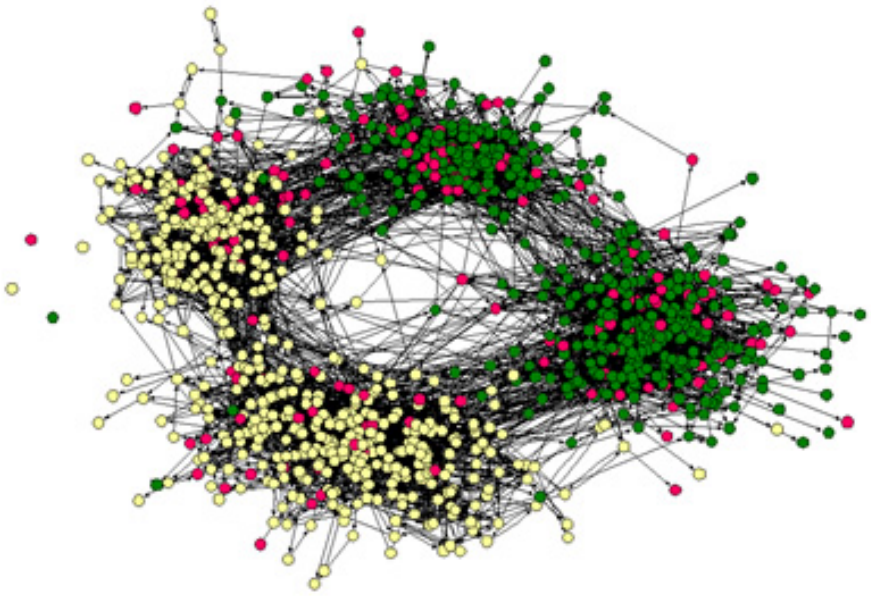

Figure 5.1: *Friendship network of children in a US school. Friendships are determined by asking the participants, and hence are directed. Student A may consider student B as their friend but not vice versa. Vertices are color coded according to race, as marked. Races: Yellow - White Race; Green - Black Race; Pink - Other [18]*

within the researched network.

Community detection is a less well posed problem than graph partitioning. Loosely stated, it is a problem of finding the natural divisions of network vertices into groups, such that there are a lot of edges between the members of the same group and a few edges between groups. However, the definition of "many" and "few" is very obscure which has lead to many debates, all proposing different definitions. Correspondingly this has lead to wide variety of different algorithms for community detection [24]. The work by Schaeffer [31] is a thorough survey about graph clustering.

5.1 Desirable community properties

In the setting of graphs, each cluster should intuitively be connected: there should be at least one path connecting each pair of vertices within a cluster. If a vertex u cannot be reached from a vertex v, they should not be placed into the same cluster. Furthermore, the paths between vertices of a cluster should be internal to the cluster. This means that in addition of the cluster C being connected in the graph G, the vertices in C should be connected by a path that only visits vertices in C. It is agreed upon that a subset of vertices is a good cluster if the induced subgraph is dense, but there a relatively few connections to vertices to the rest of the graph [10, 14, 23].

One measure for evaluation of the sparsity of connections from the cluster to the rest of the graph is the cut size $c(C, V \setminus C)$ The smaller the cut size, the better isolated the cluster is. Cluster density is determined by the density of the subgraph induced by the cluster. In literature this is referred as internal or intra-cluster density:

$$\delta_{int}(C) = \frac{|\{\{(v,u)\}|v \in C, u \in C\}|}{|C|(|C|-1)}. \tag{5.1}$$

The intra-cluster density of a given clustering of a graph G divided into k clusters C_1, C_2,...,C_k is the average of the intra-cluster densities of the included clusters:

$$\delta_{int}(G|C_1, ..., C_k) = \frac{1}{k}\sum_i \delta_{int}(C_i). \tag{5.2}$$

The external or inter-cluster density of a given clustering is defined as the ratio of intercluster edges to the maximum number of intercluster edges possible:

$$\delta_{ext}(G|C_1, ..., C_k) = \frac{|\{\{(v,u)\}|v \in C_i, u \in C_j, i \neq j\}|}{n(n-1) - \sum_l (|C_l|(|C_l|-1))}. \tag{5.3}$$

A good clustering is one whose internal density is notably higher than the density of the graph $\delta(G)$ and the intercluster density of the clustering should be considerably lower than the graph density [19, 20].

Following the above requirements of connectivity and density, the loosest definition of a graph cluster is that of a connected component, and the strictest definition is that each cluster should represent a maximal clique. Typically, interesting clustering measures tend to correspond to NP-hard decision problems. For bigger graphs, their usage is highly infeasible. In recent years, researchers have tried to introduce the concept of fuzzy clustering algorithms. It is not widely accepted approach for graph clustering, but it offers a more relaxed alternative for assigning vertices to clusters.

5.2 Measures for identifying clusters

There are two main approaches for identifying a good cluster: we can either compute some values of the vertices and then classify them into clusters based on the obtained values, or we can compute a fitness measure over the set of possible clusters and then choose among the set of possible candidates.

5.2.1 Vertex similarity

There are many clustering algorithms based on similarities between vertices. The higher the similarity between two vertices, the more inclined we are to cluster them together.

- **Distance and similarity measures**

Defining or selecting an appropriate similarity or distance function depends on the task at hand. Throughout the decades of academic research a numerous similarity measures have been used [38]. Given a data set, a distance measure should fulfill the following criteria:

1. The distance from a datum to itself is zero: $\text{dist}(d_i d_i)=0$.

2. The distances are symmetrical: dist($d_i d_j$)=dist($d_j d_i$).

3. The triangle inequality holds:

$$\text{dist}(d_i d_j) \leq \text{dist}(d_i d_k) + \text{dist}(d_k d_j) \tag{5.4}$$

Possible distance measures for two data points include Euclidean distance and the Manhattan distance. A typical example of a non-Euclidean space is that formed by vector representation of textual data. For collection of m text documents D_1, D_2,...,D_m, each term t_i that appears in at least one document is represented by a dimension. Typically non-informative words like articles and prepositions are filtered out in order to reduce the dimensionality. The total number of terms that appear in all of the associated documents is n. Each of the documents D_i is represented as a datum d_i, where the element in position j denotes the frequency at which term t_j appears in document D_i. Typically frequencies are normalized in order to eliminate the effect of document length variations. These frequencies are then multiplied by a factor that is inversely proportional to the number of documents in which the term appears, giving more weight to terms that appear in fewer documents. This product is known in the literature as term-frequency inverse-document-frequency (tf-idf). It is a practice widely used in the field of data mining [39].

Once the vectors are prepared, a similarity measure can be applied. A common measure is the cosine similarity, also known as the Ochini coefficient. This measure determines the angle between two vectors $d_i = (d_{i,1}, d_{i,2}, ..., d_{i,n})$and $d_j = (d_{j,1}, d_{j,2}, ..., d_{j,n})$ as

$$\theta(d_i, d_j) = \arccos \frac{d_i \cdot d_j}{\sqrt{\sum_k (d_{i,k}^2)}\sqrt{\sum_k (d_{j,k}^2)}} \tag{5.5}$$

The resulting measure is an angle in $[0, \pi)$. The highest similarity corresponds to the angle of zero and the most dissimilar are data whose angle is $\pi/2$. An example of using cosine similarity in clustering is the work of Lakroum et al.

- **Adjacency-based measures**

In instances where vertices lack additional properties, edges incident to the vertices can be used to derive similarity measures for the vertices, either by directly using adjacency information or through some sophisticated computation.

The simplest manner of determining whether two vertices are similar using only the adjacency information is by examining the overlap of their neighborhoods in the graph. We can compute the ratio between the intersection and the union of the sets of neighbors of two vertices, $\Gamma(v)$ and $\Gamma(u)$.

$$\omega(v,u) = \frac{|\Gamma(v) \bigcap \Gamma(u)|}{|\Gamma(v) \bigcup \Gamma(u)|} \tag{5.6}$$

is a straightforward to compute the intersection between two sets. It takes values between $[0,1]$: zero where there are no common neighbors, and one when the neighbors are identical.

- **Connectivity measures**

Clusters in graphs can be also defined through connectivity by calculating the number of different paths that exist between each pair of vertices. Vertices that belong to the same cluster should be highly connected. Edachery proposed that in a good cluster it is not absolutely necessary that two vertices are directly connected, but they should be connected by a short path. He introduces the term of threshold of the length path, which means that all vertices in the cluster should be connected by a path whose length is shorter than a chosen threshold.

5.2.2 Cluster fitness measures

Cluster fitness measures are functions that rate the quality of a given cluster or a clustering. Such measures can be used for identification of clusters, choosing between alternative clusterings and comparing different clustering algorithms.

- **Density measures**

In literature have been proposed several algorithms that search for maximal subgraphs that have a density higher than a preset threshold. Any definition of clusters as dense subgraphs is fundamentally a special case of the following problem:

Instance: Graph $G = (V, E)$, a density measure $\delta(\cdot)$ defined over vertex subsets $S \subseteq V$, a positive integer $k \leq n$, and a rational number $\xi \in [0, 1]$.

Question: Is there a subset $S \subseteq V$ such that $|S| = k$ and the density $\delta(S) \geq \xi$?

The process of maximization of cluster density is an **NP**-complete problem. In general, for large instances, approximation algorithms are a justified and feasible approach for locating dense subgraphs.

- **Cut-based measures**

To identify high-quality clusters also measures of connectivity with the rest of the graph are used. Based on cut sizes, scientists have defined measures of independence of a subgraph from the rest of the network. One of the most important measures is conductance, defined for any proper non-empty subset $S \subset V$ in graph $G = (V, E)$ as:

$$\phi(S) = \frac{c(S, V \setminus C)}{min\{deg(S), deg(V \setminus C)\}}. \tag{5.7}$$

Finding a cut with minimum conductance is a **NP**-complete problem. Variants of conductance include normalized cut, expansion, and the cut ratio.

5.3 Methods for graph clustering

Clustering methods are divided into two groups: global methods for graph clustering and local methods for graph clustering [31, 25]. In a global clustering each vertex of the input graph is assigned a cluster in the output of the method, whereas in a local clustering, the cluster assignments are only done for a certain subset of vertices, most commonly only one vertex. Below we list few of the most commonly used global methods for graph clustering.

5.3.1 Hierarchical clustering

Hierarchical clustering is not so much a single algorithm than an entire class of algorithms, with many variations and alternatives. It is an agglomerative technique in which we start with the individual vertices of a network and join the into groups.

The basic idea behind hierarchical clustering is to define a measure of similarity between vertices, based on the nature of the researched network, and then group together the vertices that are the closest or similar into common groups. This is done recursively until a quality measure converges. Similarity measures that are suitable for this kind of method for graph clustering include the cosine similarity, correlation coefficient between rows in adjacency matrix or the Euclidean distance [24].

The wide variety of available similarity measures is both strength and weakness of the hierarchical clustering method. It offers flexibility to adapt the clustering method to the specific problems, but it also means that the same method will produce different clustering results depending on the measure that has been used.

5.3.2 Divisive global clustering

Divisive clustering algorithms are a class of hierarchical methods that work top-down, recursively partitioning the graph into clusters. Each iteration splits the data into two sets, although the division could, in principle, split into more than two vertex sets.

Cuts

An intuitive approach is to search for small cuts in the network. We wish to split the graph in two by removing a cut. The removal should be done in such a way that the resulting subgraphs represent dense clusters in respect to the density of the initial graph. A well chosen cut is one that separates the graph into two or more clusters, instead of breaking into two the vertex set of any single cluster.

There are two complications with this idea. Firstly, we need to be able to make statements regarding the relative order of the subgraphs separated by a given cut.

Cutting out single vertices does not help the computation of clusters. One-by-one removal of vertices results in clusters of size one and does not reveal any higher-level structural properties. Imposition of restrictions results in the increasing the complexity of the problem, making it NP-complete.

The second complication with cut-based methods is the need to know when to stop the splitting of the graph. If we have an a priori knowledge about how the clustering should work, setting limits on the cluster order or the number of clusters can produce good results. Another approach is to optimize a chosen index of cluster quality. Hartuv and Shamir [11] propose a divisive clustering algorithm that uses density-based stopping condition. The intuition behind their decision is that vertices in the same cluster are highly connected to each other, whereas there should not be many paths leading to and from vertices from other clusters.

Spectral methods

In spectral clustering an eigenvector ar a combination of several eigenvetors is used as a vertex similarity measure for cluster computation. A comprehensive introduction to the mathematics of spectral graph clustering can be found in the book of Chung [5]. In his dissertation, McSherry [17] offers an overview of the area.

Betweenness

In order to cluster an unweighted graph $G = (V, E)$, Newman and Girvan [22] proposed to impose values of the edges based on the structural properties of the graph G. The idea is based on the vertex-betweenness proposed by Freeman [9] for sociological studies. Newman and Girvan use a betweennes of an edge (v, u) which is defined as the number of shortest paths connecting any pair of vertices in the graph that pass through the edge. Freeman previously studied the vertex-betweenness as the number of shortest paths connecting any pair of vertices in the graph that pass through a chosen vertex. It should be mentioned that there can exist numerous paths of the same shortest length that connect the same pair of vertices. Therefore each of these shortest paths should be accounted for in

proportion to their number when computing the betweenness of an edge. If there are k shortest paths connecting the pair of vertices (v, u), each of the shortest paths should have a weight of $1/k$ in the calculation of betweenness.

Newman and Girvan [10, 23] assume that edges with high betweenness are links that connect different clusters and are not intra-cluster connections. The numerous shortest paths passing through these edges are shortest paths connecting members of one cluster to those from another. The networks is split into clusters by removing the edges with highest betweenness and re-calculation of the betweenness values in the resulting network.

As mentioned before, the trick with the clustering methods based on edge removal is the decision when to stop the division. Newman [23] proposes calculation of a quality measure called modularity over the entire clustering at each iteration. The iteratve process should be stopped when there is no improvement in modularity.

Modularity is in general defined for weighted graphs, where edge weights represent some application-specific attributes. For unweighted graphs we can simply set $w(v, u) = 1$ for every edge in the network. Modularity $M(C_1, C_2, ..., C_k)$ over a specific clustering of k clusters is defined as

$$M(C_1, C_2, ..., C_k) = \sum_i \epsilon_{i,i} - \sum_{i \neq j} \epsilon_{i,j}, \tag{5.8}$$

where

$$\epsilon_{i,j} = \sum_{\{v,u\} \in E, v \in C_i, u \in C_j} w(v, u) \tag{5.9}$$

and each edge is included only once in the computation.

5.4 Louvain method

The Louvain method [2] is a simple, efficient and easy to implement method for identifying communities in large networks. It is a state-of-the-art algorithm that

was proposed by a group of scientists from Université catholique de Louvain. The method unveils hierarchies of communities in a network. Today it is one of the most widely used methods for community detection in large networks. The Louvain method is a heuristic method that is based on modularity optimization.

The method is divided in two phases that are repeated iteratively. At the beginning, each of the n vertices of the graph $G(V, E)$ are assigned to a different community. So, in this initial partition there are as many communities as there are vertices. Then, for each vertex v we consider the neighbors u of v and we evaluate the gain of modularity that would take place by removing v from its community and by placing it in the community of u. The vertex is then placed in the community for which this gain is maximum, and only if the gain is positive. If no positive gain is possible, v stays in its original community. This process is applied repeatedly and sequentially for all vertices until no further improvement can be achieved and the first phase is then complete. The first phase stops when a local maxima of the modularity is attained, that is when no individual move can improve the modularity.

The second phase of the method consists of building a new network whose vertices are now the communities found during the first phase. To do so, the weights of the links between the new vertices are given by the sum of the weight of the links between the vertices in the corresponding two new communities. Links between vertices of the same community lead to self-loops for this community in the new network.

Once the second phase is completed, then it is possible to re-apply the first phase on the newly obtained network. By construction, the number of meta-communities decreases at each pass, and as a consequence most of the computing time is used in the initial setup of the communities. The iterative passes between the first and the second phase are performed until there are no more changes and a maximum of modularity is attained. The method is reminiscent of the self-similar nature of complex networks and it naturally incorporates a notion of hierarchy.

Chapter 6

Network analysis

After the initial data collection we analyzed the collected network.

6.1 Data pre-processing

From the collected data we built the initial citation network. This network included every paper whose DOI was extracted from the processed Wikipedia articles, and their respective citation papers found on MAS. The initial network resulted in 948,791 connected by 1,539,563 edges.

As it was previously stated, we chose to collect only the papers that were citing the papers selected from Wikipedia. The result was a very sparse network, where only around 5% of the collected papers had more than one citation. Only 2.3% of all the vertices in the initial citation network were registered to have recieved 20 or more citations. The in-degree distribution can be observed on *Figure 6.1*. From the presented data on *Figure 6.1* it can be observed that the in-degree of the network seems to follow the power law.

Analyzing the out-degree distribution of the network we discovered that approximately 3% of the network vertices do not point to any other publication within the network. Only around 30% of the vertices cite two or more papers, members of the constructed network. The out-degree distribution can be observed on *Figure 6.2*. From the presented data on *Figure 6.2* it can be observed that the

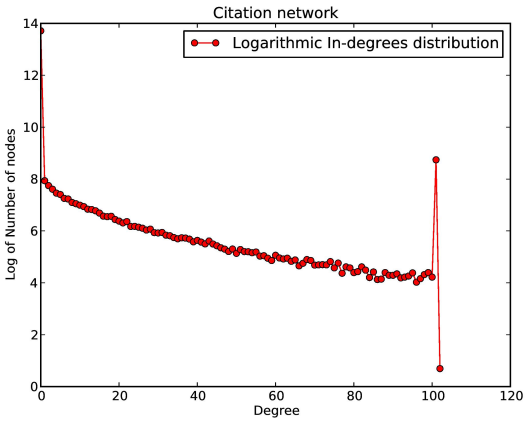

Figure 6.1: *Logarithmic representation of the in-degree distribution in the initial citation network. The straight line of the logarithmically scaled number of nodes per in-degree tends to indicate that our initial citation network follows the power law for in-degree distribution.*

out-degree of the network seems to follow the power law.

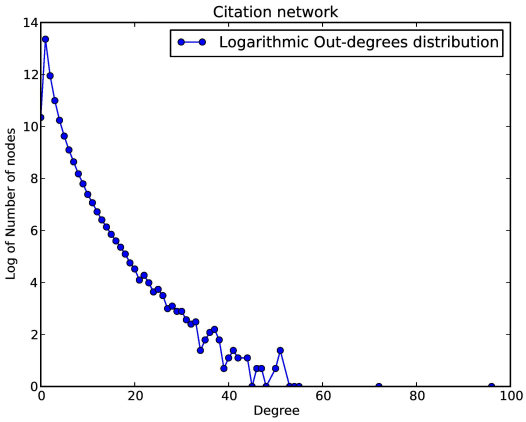

Figure 6.2: *Logarithmic representation of the out-degree distribution in the initial citation network. The straight line of the logarithmically scaled number of nodes per out-degree tends to indicate that our initial citation network follows the power law for out-degree distribution.*

Analysis of the publication year of the papers in the citation network revealed publications from 1773 to 2013, the year of data collection. MAS did not have records about the year of publication for roughly 13% of the collected publications. On *Figure 6.3* we can observe the distribution of the number of papers by year. The majority of the papers in the initial citation network were published in the last twenty years.

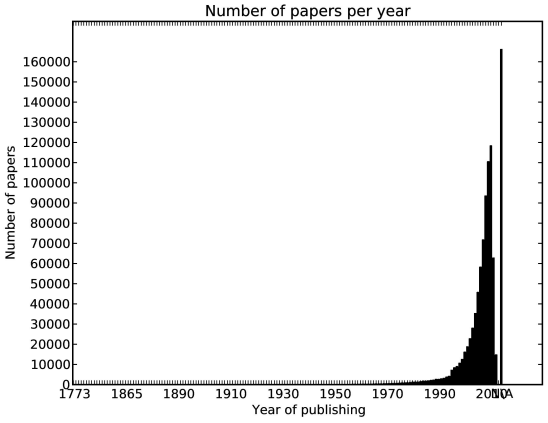

Figure 6.3: *Histogram of number of papers published by year.*

The collected information about the papers and their authors revealed that on average 5 (4.92) authors were involved in the creation of publications in the field of Psychology. Here we have to mention MAS registers maximum of 13 authors per paper. Taking into consideration this information, we can assume that the average number of authors per paper would be slightly higher. *Figure 6.4* presents the distribution of number of authors per paper.

6.2 Initial community detection

After the initial data processing, we decided to prune the network and extract only the vertices that have in-degree higher than 20. The pruning process resulted with a new graph consisting of 12,746 vertices and 7,652 edges. After removing all

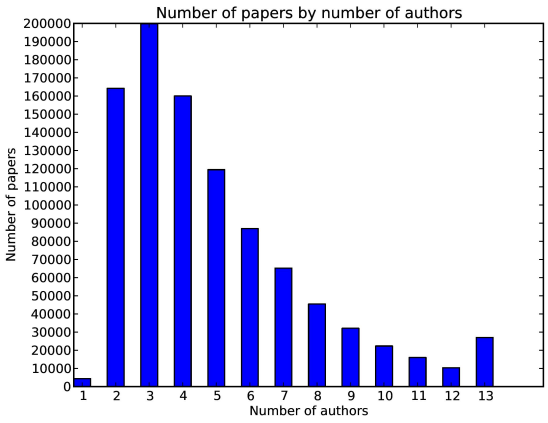

Figure 6.4: *Histogram of number of authors collaborating on a single paper. MAS registers maximum of thirteen collaborating authors per paper.*

disconnected vertices (vertices with zero in-degree and out-degree), the number of vertices was reduced to 6,329.

Interested to understand the structure of the new network, we ran the Louvain algorithm. The algorithm discovered a total of 801 communities. The initial clustering revealed that most of the vertices were connected into a big connected component, while the rest were divided into significantly smaller connected components of varying size. The smallest components consisted only of two connecected vertices. The initial division of the network into clusters can be observed on *Figure 6.5*.

Additional research of the structure of each of the components revealed that the biggest connected component included papers whose titles implied to research in the field of Psychology. Using the information about paper titles we were able to deduce that the smaller components discussed topics in the fields of Religion, formation of the Universe, Biology, etc. This should come as no surprise considering the manner in which we collected the basic papers from Wikipedia. As it was explained before, the hierarchical structure of Wikipedia articles allowed articles to be accessed from different categories. A simple example are the articles related to the psychological aspects of religion and spirituality, or the psychological ar-

ticles that explore the biological basis of behavior. We chose to extract all the papers from categories that were placed up to depth 5 from the initial Wikipedia category Psychology. This left room for the initial interdisciplinary categories that were strongly connected to Psychology to also lead the way to articles what discussed topics much distant from the ones that occupy psychologists' interest.

6.3 Community detection of Psychology papers

From the initial pruned network we extracted the vertices which were included in the largest component obtained by the application of the Louvain method. This connected component included 3,918 vertices connected by 5,732 edges.

The extracted component represented a new network, where we once again applied the Louvain method. The community detection algorithm detected 52 communities. The component division into new communities can be observed on *Figure 6.6*. The smallest community included 7 papers, while the largest community was constructed by 230 psychological publications.

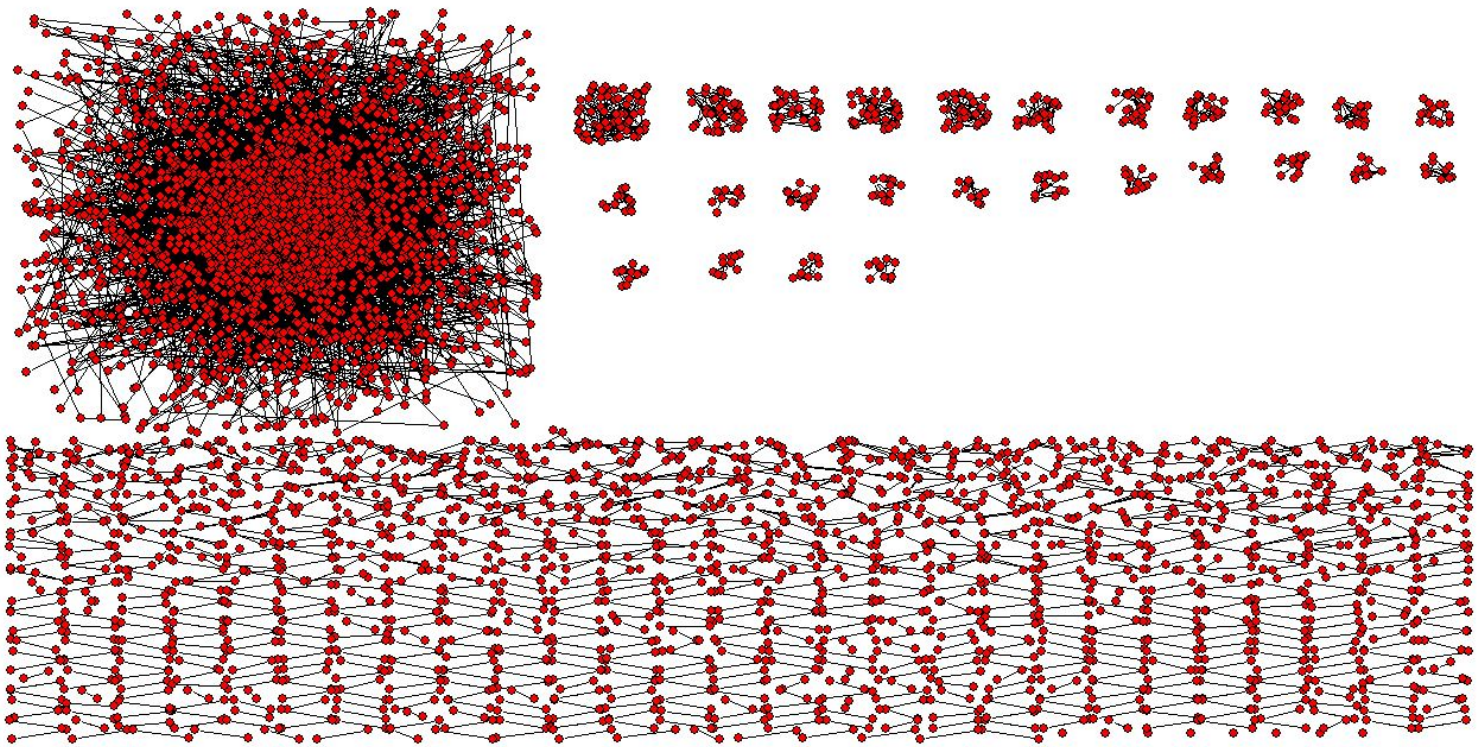

Figure 6.5: *Initial clustering of the network consisting only of papers that have collected more than 20 citations on MAS. The clustering was performed by the Louvain method which detected a total of 801 communities. Network visualization: Pajek [40]*

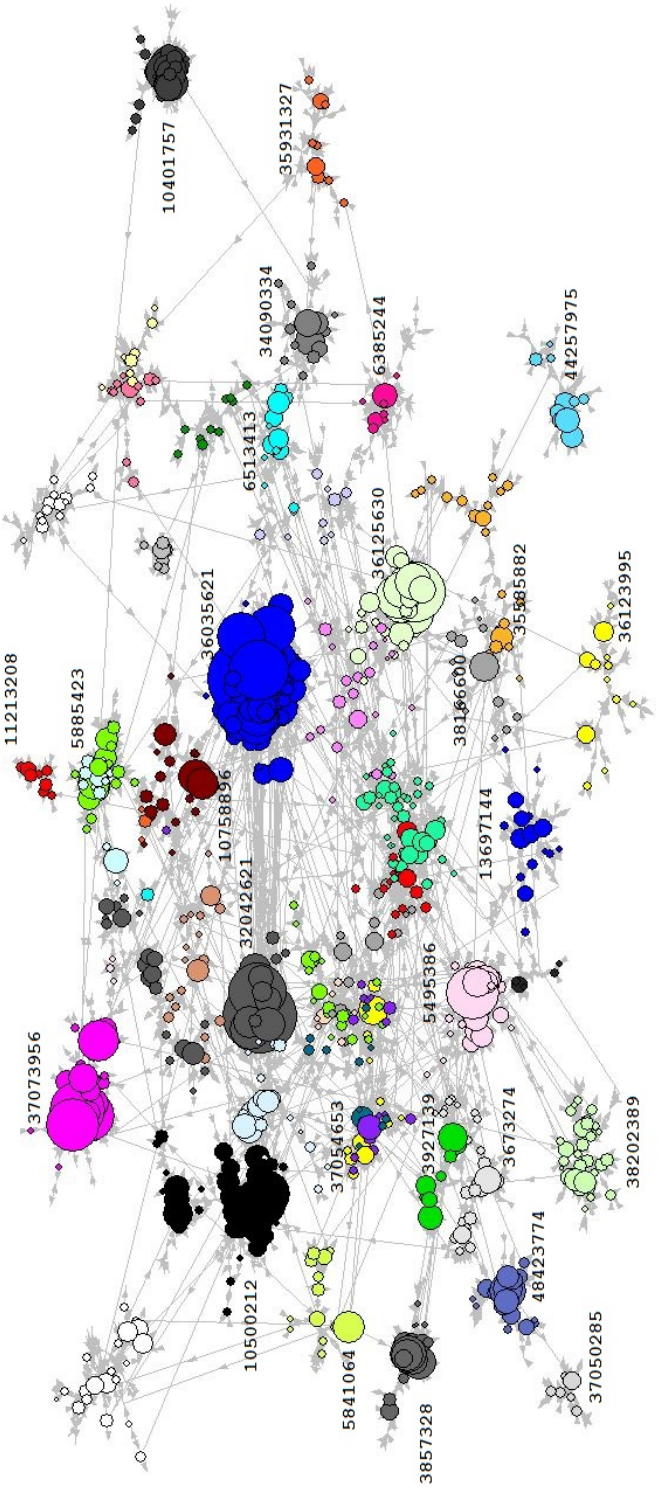

Figure 6.6: *Community detection of psychological papers. Louvain method returned a total of 52 communities. The smallest community included 7 papers, while the largest community was constructed by 230 psychological publications. Vertex size represents the betweenness value of the vertex. Papers with the highest betweenness value are labeled by their MAS ID. Networks is visualized with the Kamada-Kawai [13] visualization algorithm. Algorithm was applied twice: firstly for separation of the communities, and then for optimization of the position of vertices within their community. Table 6.1 contains additional information about each labeled vertex.*

MAS ID	Title	Year
10500212	Introduction to the Special Section on Cognitive Abilities: 100 Years After Spearman's (1904) 'General Intelligence,' Objectively Determined and Measured	2004
5841064	Brain dynamics during natural viewing conditions - A new guide for mapping connectivity in vivo	2005
3857328	The architecture of the colour centre in the human visual brain: new results and a review	2000
37054653	Prefrontal white matter volume is disproportionately larger in humans than in other primates	2005
37050285	Mutations in the α1 subunit of the inhibitory glycine receptor cause the dominant neurologic disorder, hyperekplexia	1993
48423774	Nicotinic acetylcholine receptors: from structure to brain function	NA
3673274	REMEMBERING OVER THE SHORT-TERM: The Case Against the Standard Model	2002
38202389	Infants show a facilitation effect for native language phonetic perception between 6 and 12 months	2006
3927139	The mind's best trick: how we experience conscious will	2003
5495386	The Simulating Social Mind: The Role of the Mirror Neuron System and Simulation in the Social and Communicative Deficits of Autism Spectrum Disorders	2007
13697144	Pharmacokinetics and Pharmacodynamics of Cannabinoids	2003
35585882	Neurobiology of anorexia and bulimia nervosa	2008
36123995	SAP90 Binds and Clusters Kainate Receptors Causing Incomplete Desensitization	1998
38166600	Measuring reward with the conditioned place preference (CPP) paradigm: update of the last decade	2007
44257975	Extinction of Cloudina and Namacalathus at the Precambrian-Cambrian boundary in Oman	2003
36125630	Ectopic Expression of a Microbial-Type Rhodopsin Restores Visual Responses in Mice with Photoreceptor Degeneration	2006
6385244	PRAS40 Is an Insulin-Regulated Inhibitor of the mTORC1 Protein Kinase	2007
34090334	Checking in on Cds1 (Chk2): A checkpoint kinase and tumor suppressor	2002
35931327	Nuclear export of the stress-activated protein kinase p38 mediated by its substrate MAPKAP kinase-2	1998
10401757	POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex	2004
6513413	Functional consequences of a CKIdelta mutation causing familial advanced sleep phase syndrome	2005
36035621	Autism spectrum disorders: developmental disconnection syndromes	2007
10758896	Confidence, Not Consistency, Characterizes Flashbulb Memories	2003
11213208	Cognitive Neuropsychiatric Models of Persecutory Delusions	2001
5885423	Hunter-gatherers and human evolution	2005
37073956	Genetic ancestry and the search for personalized genetic histories	2004
32042621	Neuropsychologic Functioning in Children with Autism: Further Evidence for Disordered Complex Information-Processing	2006

Table 6.1: Information about papers with highest betweenness in our citation network of psychological papers.

6.4 Naming the communities

After we divided the network component containing psychological papers into communities, we proceeded by naming the communities. The paper titles were the only available information that could be helpful in the procedure of naming the clusters. Our initial thought was to list the titles for each cluster separately, and to try and name the possible division of psychology based on the information they offered.

The average size of clusters of 75 papers, and our unfamiliarity with the research in the field of Psychology limited our capability to manually give names to the new clusters. Later we had a help from an expert psychologist, associate professor Matija Svetina, however the quantity of the available information proved too ambitious for manual manipulation.

After some consideration, we decided go back where it all started. We retraced the titles of the network vertices to the Wikipedia pages where they first originated. All of the retraced papers were connected to a DOI from the initial paper extraction from Wikipedia. Our idea was that by retracing the psychological papers to Wikipedia categories we will be able to reduce the dimensionality of the problem and consequently solve the problem of community naming. As expected, the problem dimensionality was reduced. However, the resulting dimensionality was relatively still big, and by loosing a lot of information in the process of compressing the paper titles into Wikipedia pages, it was not useful for successful identification of the extracted communities.

Finally, we decided to use the previously mentioned cosine similarity to identify the closeness of our new communities to official divisions of the research in the field of Psychology, as proposed by Professor Svetina.

6.4.1 Preparation of reference text

In order to be able to calculate the cosine similarity of the papers in the newly calculated clusters, we had to set-up a reference text. We chose the American Psychological Association (APA) division of research in the field of Psychology. APA is the largest scientific and professional organization representing psychology

in the United States. It is the world's largest association of psychologists, with more than 134,000 researchers, educators, clinicians, consultants and students as its members. APA's 54 divisions are interest groups organized by its members. Some represent sub-disciplines of psychology (e.g., experimental, social or clinical) while others focus on topical areas such as aging, ethnic minorities or trauma [http://www.apa.org/about/index.aspx].

For each of the official 54 divisions, APA enlists the scientific journal which publishes the connected papers. We gathered our reference text from MAS. The list of reference words for each of the divisions was constructed by extracting all the titles that were published in the representative journals of each division. In the cases where APA had not listed an official journal or that journal was not found on MAS, we substituted it with other scientific journal with high ranking among the scientific community and whose publications were related to the topics discussed in the division.

6.4.2 Similarity detection

Titles gathered from each of the representative journals formed a document with words relevant to the specific division. The dimensionality of the community naming problem was reduced to 52 documents. We constructed the vector space by deleting all of the uninformative words, such as articles and prepositions, and by applying the tf-idf technique to determine the importance of each word in the vector space.

After the construction of the vector space, we calculated the cosine similarity between each of the titles from the citation network of psychological papers and the reference documents for APA divisions. Each title was assigned the APA division it was most similar to. Each cluster was then left with reduced size of average 10 vertices carrying information about the relevant APA division.

6.4.3 Final community naming

Part of the APA divisions represent sub-disciplines of psychology, while the other part focuses on topical areas. Following the initial presumption that papers discussing themes in a similar topics are more likely to be connected in a citation network, we mapped each of the APA divisions into valid APA research topics. For example, the topic of Addictions was researched in the journals from three APA divisions: Psychopharmacology and Substance Abuse, Health Psychology and the Society of Addiction Psychology. Psychopharmacology and Substance Abuse promotes teaching, research and dissemination of information regarding the effects of drugs on behavior. Health Psychology seeks to advance contributions of psychology to the understanding of health and illness through basic and clinical research, education and service activities and encourages the integration of biomedical information about health and illness with current psychological knowledge. Society of Addiction Psychology promotes advances in research, professional training, and clinical practice within the broad range of addictive behaviors including problematic use of alcohol, nicotine and other drugs and disorders involving gambling, eating, sexual behavior or spending. At the time of writing there were 85 active research APA topics.

The resulting community naming can be observed in *Figure 6.7*. It is clear from the visualized graph that papers with the highest betweenness value belong to the divisions: Intellectual and Developmental Disabilities, Society for Child and Family Policy and Practice, and Society of Clinical Psychology. Papers from these divisions act as connectors for papers published in other divisions. The researched showed that the most explored topics by Psychologists were the topics about children, women, aging and ethnic minorities. This should not be surprising, since the way today's children are shaped affects our future, and women, especially mothers, are instrumental part of this shaping process. Aging is another topic that is very popular. Psychologists try to explain the whole process from psychological point of view, and offer guidelines for its acceptance. In today's globalization, an interesting topic for research are also the ethnic minorities, and their treatment and inclusion into the society.

In *Figure 6.8* we can observe the network of our final 52 clusters. Edge weight represents scaled representation of the number of edges pointing to or from a chosen cluster. The network is strongly connected. We would like to believe that this is due to the nature of Psychology, and the need of papers from different divisions to cite each other. As can be seen from the network, the resulting edges are rational. For example, papers from Intellectual and Developmental Disabilities cite very often publications from Society for Child and Family Policy and Practice. Many times the intellectual and developmental disabilities in a child are as a result of long term addiction. This is also implied by the connection between divisions Intellectual and Developmental Disabilities and Society of Addiction Psychology. Behavioral Neuroscience and Comparative Psychology and Trauma Psychology are two topics whose researches inevitably cross paths.

We need to mention that the naming process was done with our own best judgment, in accordance with the results for closest divisions and topics obtained with the cosine similarity measure. It is highly possible that the resulting naming is subjective and not accurate. Since currently there are no measures to confirm the quality of our communities, we have to be a bit skeptical about the network division too.

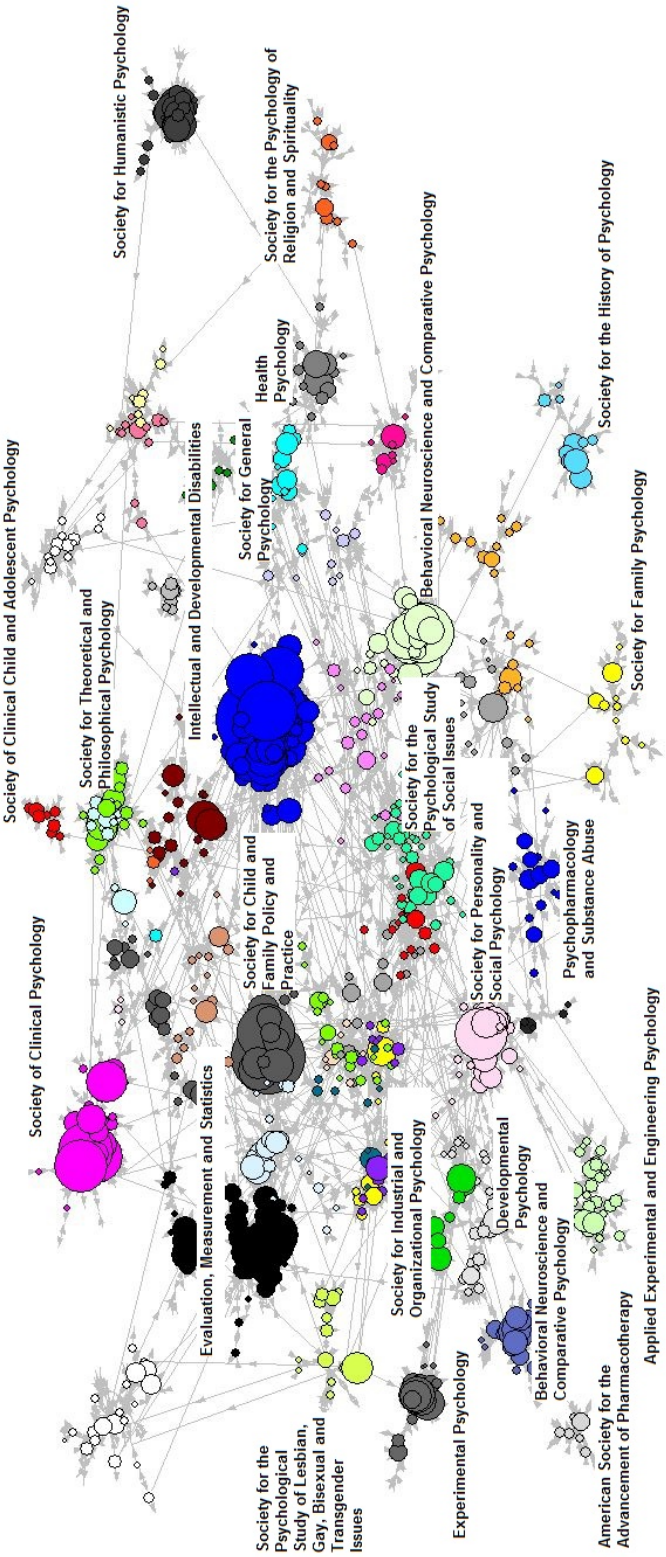

Figure 6.7: *Named communities of our network of psychological papers.*

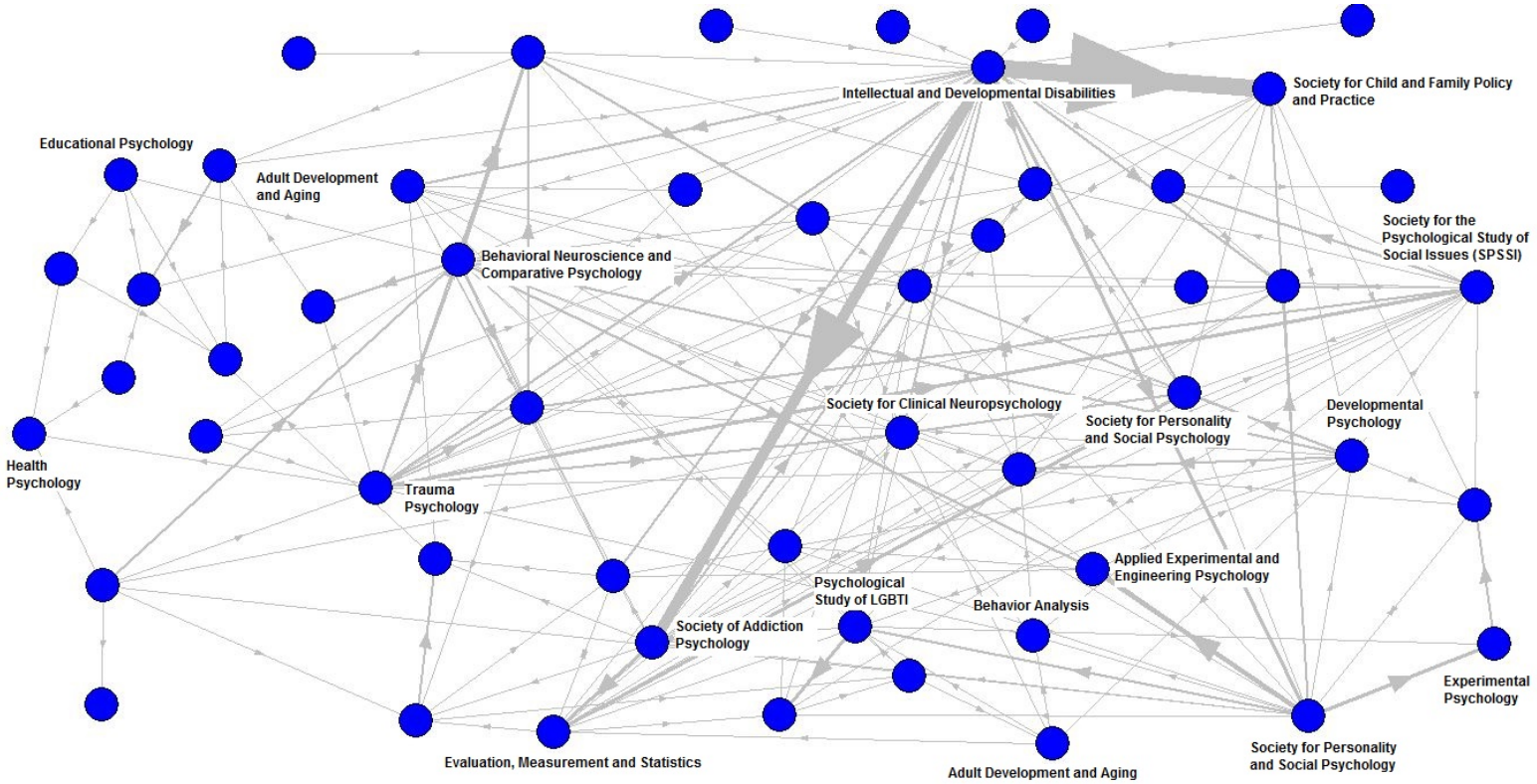

Figure 6.8: *Network of clusters. Weighted directed edges from cluster A to cluster B represent a scaled number of papers belonging to cluster A that cite papers members of cluster B.*

Chapter 7

Conclusion

In this thesis we explored the possibility of extraction of research disciplines from Psychology based only on the citation network formed by papers published in the field. We were interested to see whether we could differentiate between various Psychology disciplines and topics based only on the structure of the citation network.

The unavailability of central database for publications in the field of Psychology and their citation demanded finding an alternative way for construction of the citation network. Encouraged by the structure of Wikipedia and their organization of articles, we decided to perform a depth search for psychology publications in Wikipedia categories and articles connected to the principal Wikipedia category of Psychology. The inconsistencies in citation styles meant that we had to adapt our extraction process, and only extract publications with referenced DOI. This limitation affected the end result citation network and it almost certainly had implications on the final result and the detected communities.

After the initial DOI extraction, we constructed the citation network with the extracted publications from Wikipedia and their respective citing papers on MAS. The building process was performed only one level deep, meaning the acquisition of new papers was halted when the first layer of citing papers was collected. In a way, this constructed a one way street, with directed edges from MAS citing papers to initial cited papers from Wikipedia. However, the quantity of collected

data and its nature meant formation of a network which expressed the qualities reflected by most of the real-world networks.

The application of state-of-the-art algorithm for community detection on the initial citation network revealed one big connected component and a numerous unconnected small communities. Upon closer inspection, we concluded that the sizable component was constructed by papers in the field of Psychology, while the other unconnected components revealed publications from other unconnected fields which could initially be reached from our depth search of the Wikipedia articles structure.

Our analysis of the psychological components detected 52 communities, a number close to the official 54 divisions of Psychology introduced by the American Psychological Association. The dimensionality of the problem and our inexperience with the research and the scientific terms in the field of Psychology proved to be a handicap in the attempt to rate the division process and to name the resulting communities. We used the measure of cosine similarity, often used for classification of textual data, to facilitate the process of naming the clusters. The obtained results for relevant topics and divisions simplified our decision for assignment of appropriate community names.

The resulting network of interconnected communities, and their logical links can provide a solid ground for conclusion that our approach for subdiscipline detection in Psychology was partly successful. We must not forget that the naming process was also affected by our subjective interpretation of the results obtained by the applied cosine similarity.

The data obtained through the process of data collection and the data analysis offer a great platform for further improvement and additional research of the citation networks of scientific disciplines. The first thing that comes to mind is refinement of the collecting process of psychological papers. Instead of using Wikipedia categories connected to Psychology, we can start with collecting publications from the most influential psychological journals with records on MAS. This hopefully will provide us with a complete initial database. Collection of psychological papers from the most influential journals in the field will provide us with a

compact year by year review of publications and research development in different disciplines of the science. Further, we can extend our additional paper acquisition to papers which were also cited by the papers in the initial database, or collect papers which are connected with the initially collected papers to a certain depth. For example, papers which cite papers citing the initially collected publications, hold connections with depth two with the papers from the initial database.

Further investigation in the pattern properties of the constructed citation network may reveal new measures for improved community detection. This and a more sophisticated method for similarity measurement of textual data may lead to a more objective process of community naming. The procedures and results listed in this thesis can provide a good basis for future further research of citation network with the intention of detecting valid research disciplines. It is our opinion that it will be an interesting reading to re-apply the same ideas of discipline detection to other sciences.

Appendix A

Used abbreviations and symbols

	Description
APA	American Psychological Association
DOI	Digital Object Identifier
GSC	Google Scholar
MAS	Microsoft Academic Search
NP	Non-deterministic Polynomial-time
N	Network
G	Graph
V	Set of vertices
E	Set of edges
M	Modularity
n	Number of vertices
v_i	i-th vertex
e_i	i-th edge
C_i	i-th cluster

Bibliography

[1] A. D. Anastasiadis, M. P. De Albuquerque, M. P. De Albuquerque, D. B. Mussi. "Tsallis q-exponential describes the distribution of scientific citations - A new characterization of the impact", *Scientometrics* , vol. 83, no. 1, pp. 205–218, 2008.

[2] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre. "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics-Theory and Experiment*, vol. 10, no. 10, 2008.

[3] S. Brin, L. Page. "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, vol. 30, pp. 107–117, 1998.

[4] P. Chen, H. Xieb, S. Maslovc, S. Rednera. "Finding scientific gems with Google's PageRank algorithm", *Journal of Informetrics*, vol. 1, 2007.

[5] F. R. K. Chung. *Spectral Graph Theory*, American Mathematical Society, 1997.

[6] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification*, 2nd ed., John Wiley & Sons, Inc., New York, NY, USA, 2001.

[7] P. Erdős, A. Rényi. "On random graphs I", *Publicationes Mathematicae*, no. 6, pp. 290–297, 1959.

[8] P. Erdős, A. Rényi. "On the evolution of random graphs", *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, no. 5, pp. 10–17, 1960.

[9] L. C. Freeman. "A set of measures of centrality based upon betweenness", *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[10] M. Girvan, M. E. J. Newman. "Community structure in social and biological networks" , *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[11] E. Hartuv, R. Shamir. "A clustering algorithm based on graph connectivity", *Information Processing Letters*, vol. 76, pp. 175–181, 2000.

[12] P. Jacsó. "The pros and cons of Microsoft Academic Search from a bibliometric perspective", *Online Information Review*, vol. 35, no. 6, pp. 983–997, 2011.

[13] T. Kamada, S. Kawai. "An algorithm for drawing general undirected graphs", *Information Processing Letters*, vol. 31, no. 1, pp. 7–15, 1989.

[14] J. Kleinberg, S. Lawrence. "The Structure of the Web", *Science*,vol. 294, pp. 1849–1850, 2001.

[15] J. Kleinberg, E. Tardos. "Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields", *Journal of the ACM*, vol. 49, no. 5, pp. 616–639, 2002.

[16] N. D. Martinez. "Constant connectance in community food webs", *American Naturalist*, vol. 139, pp. 1208–1218, 1992.

[17] F. McSherry. "Spectral Methods for Data Analysis", Ph.D Thesis, University of Washington, Seattle, WA, USA, 2004.

[18] J. Moody. "Race, school integration, and friendship segregation in America", *American Journal of Sociology* vol. 107, no. 3, pp. 679–716, 2001.

[19] M. E. J. Newman. "Detecting community structure in networks", *The European Physical Journal B*, vol. 38, no. 2 pp. 321–330, 2004.

[20] M. E. J. Newman. "Finding community structure in networks using the eigenvectors of matrices", *Physical Review E*, vol. 74, no. 3, 2006.

[21] M. E. J. Newman. "The structure and function of complex networks", *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[22] M. E. J. Newman, M. Girvan. "Mixing Patterns and Community Structure in Networks", 2003.

[23] M. E. J. Newman. "Fast algorithm for detecting community structure in networks", *Physical Review E*, vol. 69, no. 6, 2004.

[24] M. E. J. Newman. *Networks: An Introduction*, Oxford University Press Inc., New York, 2010.

[25] M. E. J. Newman, M. Girvan. "Finding and evaluating community structure in networks", *Physical Review E*, vol. 69, no. 2, 2004.

[26] J. L. Ortega, I. F. Aguillo. "Microsoft Academic Search and Google Scholar Citations: A comparative analysis of author profiles", *Journal of the Association for Information Science and Technology*, Feb. 26th, 2014.

[27] D. J. de Solla Price. "Networks of Scientific Papers", *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[28] F. Radicchi, S. Fortunato, C. Castellano. "Universality of citation distributions: Toward an objective measure of scientific impact", *Proceedings of The National Academy of Sciences*, vol. 105, no. 45, pp. 17268–17272, 2008.

[29] S. Redner. "How popular is your paper? An empirical study of the citation distribution", *European Physical Journal B*, vol. 4, no. 2, pp. 131–134, 1998.

[30] S. Redner. "Citation Statistics from 110 Years of Physical Review", *Physics Today*, vol. 58, no. 6, pp. 49–54, 2005.

[31] S. E. Schaefefer. "Graph clustering", *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.

[32] P. O. Seglen. "The Skewness of Science", *Journal of The American Society for Information Science and Technology*, vol. 43, no. 9, pp. 628–638, 1992.

[33] R. Solomonoff, A. Rapoport. "Connectivity of random nets", *The Bulletin of mathematical biophysics*, vol. 13, pp. 107–117, 1951.

[34] A. F. J. van Raan. "Two-step competition process leads to quasi power-law income distributions - Application to scientific publication and citation distributions", *Physica A-statistical Mechanics and Its Applications*, vol. 298, pp. 530–536, 2001.

[35] D. Walker, H. Xie, K. K. Yan, S. Maslov. "Ranking scientific publications using a model of network traffic", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 6, 2007.

[36] M. L. Wallace, V. Larivière, Y. Gingras. "Modeling a century of citation distributions", *Journal of Informetrics*, vol. 3, no. 4, pp. 296–303, 2009.

[37] D. J. Watts, S. H. Strogatz. "Collective dynamics of 'small-world' networks", *Nature*, vol. 393, pp. 440–442, 1998.

[38] W. T. Williams, M. B. Dale, P. Macnaughton-Smith. "An objective method of weighting in similarity analysis", *Nature*, vol. 201, no. 4917, pp. 426–426, 1964.

[39] S. M. Wong, Y. Y. Yao. "An information-theoretic measure of term specificity", *Journal of the American Society for Information Science*, vol. 43,no. 1, pp. 54–61, 1992.

[40] Pajek, available on: `http://pajek.imfm.si/doku.php?id=download`.

[41] Microsoft Academic: Windows Azure Marketplace, available on: `http://datamarket.azure.com/dataset/mrc/microsoftacademic`.