PUC

ISSN 0103-9741

Monografias em Ciência da Computação

n° 20/10

Social Media Savvy: exploiting Orkut data

Dárlinton Barbosa Feres Carvalho

Carlos José Pereira de Lucena

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO

RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22453-900

RIO DE JANEIRO - BRASIL

Monografias em Ciência da Computação, No. 20/10 ISSN: 0103-9741
Editor: Prof. Carlos José Pereira de Lucena December 2010

Social Media Savvy: exploiting Orkut data

Dárlinton Barbosa Feres Carvalho, Carlos José Pereira de Lucena

darlinton@acm.org, lucena@inf.puc-rio.br

Abstract.

This work presents a new method to savvy the social media from Orkut. The focus is on the user interaction in the communities. The social media is primarily available in the conversation present on the community forum. However, the community association has also a special meaning in the system, because users make use of the associations through badges defining their profiles. The method creates a model of community relationships to extract intriguing relationships of the social data. Special plots for data visualization help the social media expert to analyze the data. Three case studies illustrate the application of the proposed method, and present considerations about it as well.

Keywords: Social Media, Online Social Networks, Orkut, Brand Monitoring, Modeling and Measurement Social Media.

Resumo.

Este trabalho apresenta um novo método para auxiliar no entendimento de mídias sociais no Orkut. O foco da análise é na interaçãoo entre os usuários dentro das comunidades no sistema. A mídia social está disponível primeiramente na conversação presente no fórum da comunidade em análise. Entretanto, a associação do usuário a uma comunidade também tem um significado especial, porque os usuários usam essas associações como meio para definir seus perfis. O método proposto cria um modelo de relacionamento entre comunidades para gerar relacionamentos dos dados sociais. Gráficos especiais para a visualização dos dados ajudam os especialistas em mídias sociais a analisarem os dados. Três estudos de caso ilustram a aplicação do método proposto e também apresentam considerações sobre sua utilização.

Palavras-chave: Mídia Social, Redes Sociais Online, Orkut, Monitoramento de Marcas, Modelagem e Mensuração de Mídia Social.

In charge of publications:
Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC-Rio Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22451-900 Rio de Janeiro RJ Brasil
Tel. +55 21 3527-1516 Fax: +55 21 3527-1530
E-mail: bib-di@inf.puc-rio.br
Web site: http://bib-di.inf.puc-rio.br/techreports/

Contents

1 Introduction

Social media is content generated by people in social contexts. Web 2.0 has enabled users to publish content on the Internet in various sorts of ways (e.g. blogs, photos, notes on social networks sites). With more than two billion users[1], the Internet is the mainstream channel to study the social media in our society. In Brazil, Orkut[2] is the main social media source; it is the leader regarding number of users and traffic[3]. This work presents a new method to savvy the social media from the Orkut system.

As a new trend on the Internet, there are many works on the recent literature about social media. The mainstream theory used to analyze social networks is based on complex networks studies (da F. Costa, Rodrigues, Travieso & Boas 2007). There are studies trying to characterize the population of sites on the Internet, like Facebook (Nazir, Raza & Chuah 2008), Twitter (Huberman, Romero & Wu 2008) and YouTube (Cheng, Dale & Liu 2008). Other studies aim to understand the users on social networks, special users on these networks, the network evolution and its implications (Benevenuto, Rodrigues, Cha & Almeida 2009, Wilson, Boe, Sala, Puttaswamy & Zhao 2009, Ahmed, Berchmans, Neville & Kompella 2010, Bigonha, Cardoso, Moro, Almeida & Goncalves 2010). Recent studies are questioning the overrating of the social network studies and even questioning if the systems are general-purpose social network sites or only a news media site with social networks support (Kwak, Lee, Park & Moon 2010, Cormode, Krishnamurthy & Willinger 2010). The market research area is also looking for better tools to catch up with the new trend[4]. This work sheds light on the Orkut social media, defining a new method to help analyze the data.

The method proposed in this work relies mainly on the Orkut communities. Most of methods try to describe social data from the social network point of view, but this work focus on the user interaction with the communities. The social media is available in the conversation present on the community forum, as posts organized by topic. In Orkut, the community membership also has a special meaning. Orkut users make use of the community associations through badges defining their profiles. Based on this fact, the method relies on a model of communities relationships to extract intriguing relationships of the social data.

The case study illustrates the proposed method applied to two different contexts. The first one aims at the analysis of Brazilian elections, looking at one candidate community around 80 thousand users. The first context is a good example of the efforts necessary to execute the method. The second context is the analysis of brands, with two different case studies. It has a comprehensive visualization analysis.

The rest of this paper is organized as follow. The methodology, with the epistemology considered on the method creation, is presented in Section 2. Section 3 has three case studies showing the application of the method. The conclusions about this work are found in Section 4. Extra visualization plots for the case studies are available in the Appendix.

[1] http://www.internetworldstats.com/stats.htm

[2] http://www.orkut.com

[3] http://www.comscore.com/Press_Events/Press_Releases/2010/10/Orkut_Continues_to_Lead_Brazil_s_Social_Networking_Market_Facebook_Audience_Grows_Fivefold

[4] http://blogs.forrester.com/zach_hofer_shall/10-07-12-forrester_wave_listening_platforms_q3_2010

2 Methodology

The study of social media has its roots in social science. However, the use of modern communication technology such as the Internet (i.e. Social Web, Web 2.0 apps) brings a new perspective for the study of social media. Following this paradigm shift, this work introduces a new method to exploit the data available on the Web. This section presents the epistemology considered on the methodology.

2.1 Research Questions

As an exploratory work, it is defined research questions in a broad spectrum. The questions follow in a progressive order as the classic scientific studies (data gathering, modeling, and analysis). The answers define a new method to study social media in the defined context.

Q1) What is the useful information that can be extracted from Orkut?

Q2) How to represent the relationship between the subject of analysis and the extracted data?

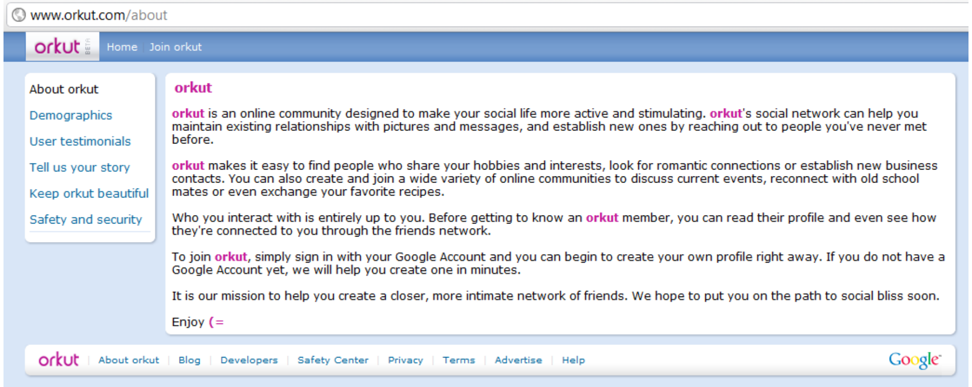
Q3) What is a good sampling size?

Q4) What kind of analysis can be done from the model?

2.2 Orkut Social Media

The foundation of the method proposed in this work is on the main concepts and features available on the Orkut system. The aim is to exploit the public social media. This section explains the system and give hints to answering the first research question.

Orkut (`http://www.orkut.com/`)is an all-purpose online community site designed to facilitate social network interactions. The social network amenities includes sharing pictures and messages and also an organized way to maintain existing relationships, retrieving lost connections and even establishing new ones. The site belongs to Google and use is free. The Figure 1 shows the official description of the site.



orkut

orkut is an online community designed to make your social life more active and stimulating. orkut's social network can help you maintain existing relationships with pictures and messages, and establish new ones by reaching out to people you've never met before.

orkut makes it easy to find people who share your hobbies and interests, look for romantic connections or establish new business contacts. You can also create and join a wide variety of online communities to discuss current events, reconnect with old school mates or even exchange your favorite recipes.

Who you interact with is entirely up to you. Before getting to know an orkut member, you can read their profile and even see how they're connected to you through the friends network.

To join orkut, simply sign in with your Google Account and you can begin to create your own profile right away. If you do not have a Google Account yet, we will help you create one in minutes.

It is our mission to help you create a closer, more intimate network of friends. We hope to put you on the path to social bliss soon.

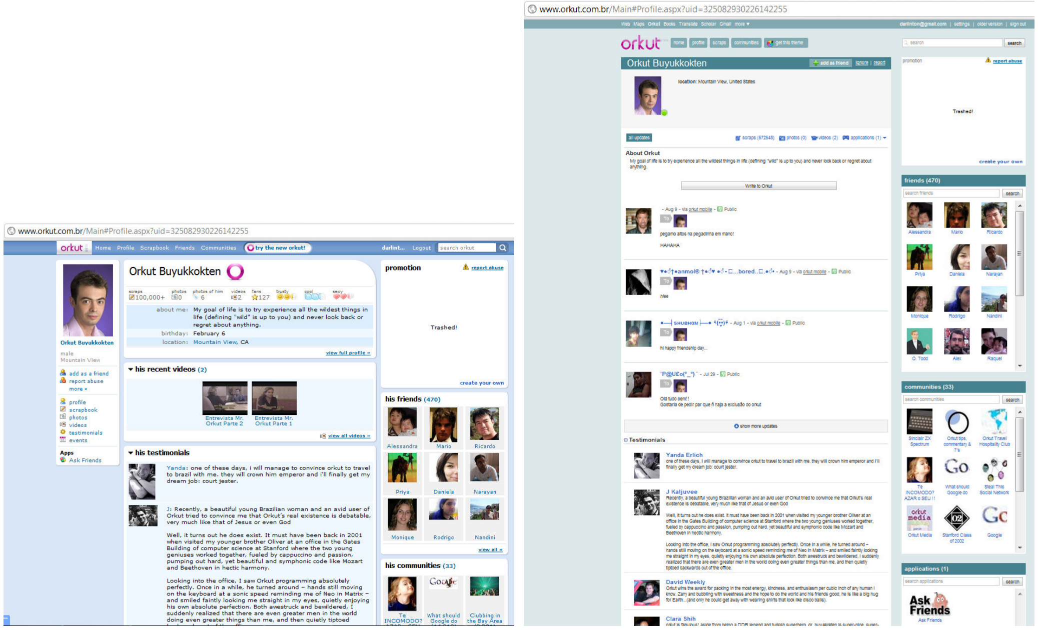
Enjoy (=

Figure 1: Orkut about

There are more than 100 million active users on Orkut[5]. Most users are Brazilians. The site has demographic information available at `http://www.orkut.com/MembersAll`.

[5] `http://www.alexa.com/siteinfo/orkut.com`, `http://en.wikipedia.org/wiki/Orkut`

However, the site does not provide more information about its use or even an API to access the available data.



about me: My goal of life is to try experience all the wildest things in life (defining "wild" is up to you) and never look back or regret about anything.
birthday: February 6
location: Mountain View, CA

Yanda: one of these days, i will manage to convince orkut to travel to brazil with me. they will crown him emperor and i'll finally get my dream job: court jester.

J: Recently, a beautiful young Brazilian woman and an avid user of Orkut tried to convince me that Orkut's real existence is debatable, very much like that of Jesus or even God

Well, it turns out he does exist. It must have been back in 2001 when visited my younger brother Oliver at an office in the Gates Building of computer science at Stanford where the two young geniuses worked together, fueled by cappuccino and passion, pumping out hard, yet beautiful and symphonic code like Mozart and Beethoven in hectic harmony.

Looking into the office, I saw Orkut programming absolutely perfectly. Once in a while, he turned around – hands still moving on the keyboard at a sonic speed reminding me of Neo in Matrix – and smiled faintly looking me straight in my eyes, quietly enjoying his own absolute perfection. Both awestruck and bewildered, I suddenly realized that there are even greater men in the world doing even greater things than me, and then quietly tiptoed

location: Mountain View, United States

About Orkut
My goal of life is to try experience all the wildest things in life (defining "wild" is up to you) and never look back or regret about anything.

- Aug 9 - via orkut mobile - Public
To
pegamo altos na pegadinha em mano!
HAHAHA

♥•♂†•anmol® †•♂♥ •♂•□...bored..□.•♂• - Aug 9 - via orkut mobile - Public
To
hiee

•—┤SHUBHαм├—• ٩(͡•͡•)۶ - Aug 1 - via orkut mobile - Public
To
hi happy friendship day...

¨P@U£o(°_°)¨ - Jul 29 - Public
To
Olá tudo bem!!
Gostaria de pedir par que ñ haja a exclusão do orkut

Yanda Erlich
one of these days, I will manage to convince orkut to travel to brazil with me. they will crown him emperor and I'll finally get my dream job: court jester.

J Kaljuvee
Recently, a beautiful young Brazilian woman and an avid user of Orkut tried to convince me that Orkut's real existence is debatable, very much like that of Jesus or even God

Well, it turns out he does exist. It must have been back in 2001 when visited my younger brother Oliver at an office in the Gates Building of computer science at Stanford where the two young geniuses worked together, fueled by cappuccino and passion, pumping out hard, yet beautiful and symphonic code like Mozart and Beethoven in hectic harmony.

Looking into the office, I saw Orkut programming absolutely perfectly. Once in a while, he turned around – hands still moving on the keyboard at a sonic speed reminding me of Neo in Matrix – and smiled faintly looking me straight in my eyes, quietly enjoying his own absolute perfection. Both awestruck and bewildered, I suddenly realized that there are even greater men in the world doing even greater things than me, and then quietly tiptoed backwards out of the office.

David Weekly
Orkut wins the award for packing in the most energy, kindness, and enthusiasm per cubic inch of any human I know. Zany and bubbling with sweetness and the hope to do the world and his friends good, he is like a big hug for Earth...(and only he could get away with wearing shirts that look like disco balls).

Clara Shih
orkut is fabulous! aside from being a DDR legend and turkish superhero, dr. buyukkokten is super-nice, super-

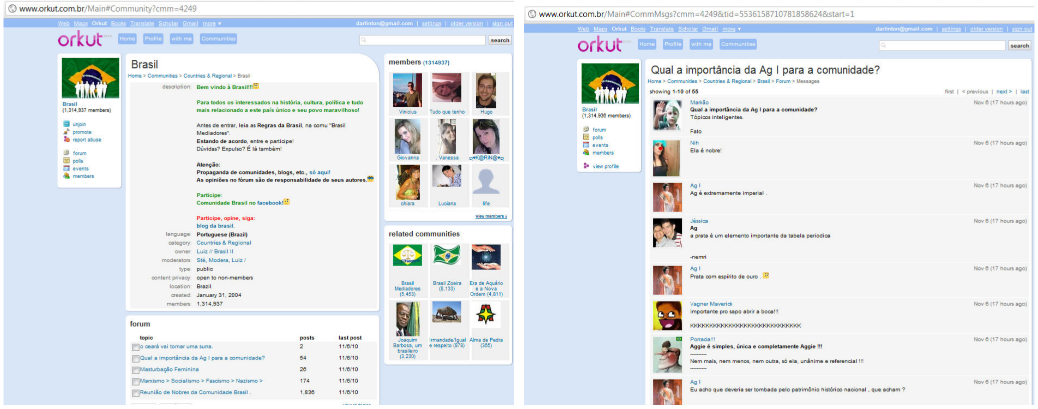
Figure 2: Orkut Home Page – Old (left) and Actual (right) Version

The first-class elements on Orkut System are the communities and users. The system has improved since its debut, but the changes are adding of new features. The actual and the old design of the main home page can be seen in the Figure 2.

A community in Orkut has an owner, description, users and a forum. Along the system enhancements, the communities have gotten new features such as polls, event announcements, suggestions of related communities, moderators, etc. A preview of the community interface and its forum is available at Figure 3. The forum is simple and organized by topics. The users can post messages on a topic, or even create a new one to start a new topic discussion. There is a simple privacy control of the data (open/public or hidden/members only), open being the default disclosure mode.

In the Orkut System, a user has a profile, friends and communities. The privacy control of the user information is mainly about the profile data, being the social network and communities available with public disclosure (always). Figure 4 presents the profile of the system creator in the old design version. Although, the last redesign of the site profile does not have the general information about the user, having now only contact data (i.e. address, phone, e-mail). The profile configuration screen with all the data available on the new design is shown in Figure 5. Regarding the profile, the users found an easier, more flexible and unexpected way to define themselves. The users join communities to define themselves, seeing the communities on their profiles like badges. The communication among the users is evolving and the privacy control is getting better for all methods of interaction as well. Because of the privacy issues, this work focuses only on the publicly disclosed data.

The system offers a comprehensive search feature. Figure 6 shows the Universal Search Page. The search focus is on users, communities, topics or everything together. There


www.orkut.com.br/Main#Community?cmm=4249
Web Maps Orkut Books Translate Scholar Gmail more ▾
darlinton@gmail.com | settings | older version | sign out
orkut
Home
Profile
with me
Communities
search
unjoin
promote
report abuse
forum
polls
events
members
view members »
view all topics

Brasil
(1,314,937 members)

Brasil

Home > Communities > Countries & Regional > Brasil

description: Bem vindo à Brasil!!!

Para todos os interessados na história, cultura, política e tudo mais relacionado a este país único e seu povo maravilhoso!

Antes de entrar, leia as Regras da Brasil, na comu "Brasil Mediadores".
Estando de acordo, entre e participe!
Dúvidas? Expulso? É lá também!

Atenção:
Propaganda de comunidades, blogs, etc., só aqui!
As opiniões no fórum são de responsabilidade de seus autores.

Participe:
Comunidade Brasil no facebook!

Participe, opine, siga:
blog da brasil.

language:	Portuguese (Brazil)
category:	Countries & Regional
owner:	Luiz // Brasil II
moderators:	Sté, Modera, Luiz /
type:	public
content privacy:	open to non-members
location:	Brazil
created:	January 31, 2004
members:	1,314,937

members (1314937)

Vinicius, Tudo que tenho, Hugo, Giovanna, . Vanessa, ღ♥K@RiN@♥ღ, chiara, Luciana, life

related communities

Brasil Mediadores (5,453); Brasil Zoeira (8,133); Era de Aquário e a Nova Ordem (4,811); Joaquim Barbosa, um brasileiro (3,230); Irmandade/Igual e respeito (878); Alma de Pedra (365)

forum

topic	posts	last post
o ceará vai tomar uma surra.	2	11/6/10
Qual a importância da Ag I para a comunidade?	54	11/6/10
Masturbação Feminina	26	11/6/10
Marxismo > Socialismo > Fascismo > Nazismo >	174	11/6/10
Reunião de Nobres da Comunidade Brasil .	1,836	11/6/10

www.orkut.com.br/Main#CommMsgs?cmm=4249&tid=5536158710781858624&start=1

Brasil
(1,314,936 members)

Qual a importância da Ag I para a comunidade?

Home > Communities > Countries & Regional > Brasil > Forum > Messages

showing 1-10 of 55 — first | < previous | next > | last

Markão — Nov 6 (17 hours ago)
Qual a importância da Ag I para a comunidade?
Tópicos inteligentes.

Fato

Nih — Nov 6 (17 hours ago)
Ela é nobre!

Ag I — Nov 6 (17 hours ago)
Ag é extremamente imperial .

Jéssica — Nov 6 (17 hours ago)
Ag
a prata é um elemento importante da tabela periodica

-nemri

Ag I — Nov 6 (17 hours ago)
Prata com espírito de ouro .

Vagner Maverick — Nov 6 (17 hours ago)
importante pro sapo abrir a boca!!!

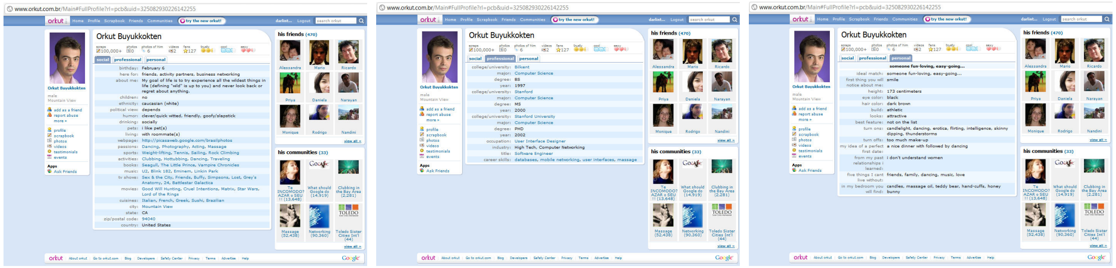
KKKKKKKKKKKKKKKKKKKKKKKKKKKKK

Porrada!!! — Nov 6 (17 hours ago)
Aggie é simples, única e completamente Aggie !!!

Nem mais, nem menos, nem outra, só ela, unânime e referencial !!!

Ag I — Nov 6 (17 hours ago)
Eu acho que deveria ser tombada pelo patrimônio histórico nacional , que acham ?

Figure 3: Orkut Community Page – Main View (left) and Forum (right)



Orkut Buyukkokten
male
Mountain View

Orkut Buyukkokten

scraps 100,000+ photos 0 photos of him 6 videos 2 fans 127 trusty cool sexy

birthday:	February 6
here for:	friends, activity partners, business networking
about me:	My goal of life is to try experience all the wildest things in life (defining "wild" is up to you) and never look back or regret about anything.
children:	no
ethnicity:	caucasian (white)
political view:	depends
humor:	clever/quick witted, friendly, goofy/slapstick
drinking:	socially
pets:	i like pet(s)
living:	with roommate(s)
webpage:	http://picasaweb.google.com/brasilphotos
passions:	Dancing, Photography, Acting, Massage
sports:	Weight-lifting, Tennis, Sailing, Rock Climbing
activities:	Clubbing, Hottubbing, Dancing, Traveling
books:	Seagull, The Little Prince, Vampire Chronicles
music:	U2, Blink 182, Eminem, Linkin Park
tv shows:	Sex & the City, Friends, Buffy, Simpsons, Lost, Grey's Anatomy, 24, Battlestar Galactica
movies:	Good Will Hunting, Cruel Intentions, Matrix, Star Wars, Lord of the Rings
cuisines:	Italian, French, Greek, Sushi, Brazilian
city:	Mountain View
state:	CA
zip/postal code:	94040
country:	United States

college/university:	Bilkent
major:	Computer Science
degree:	BS
year:	1997
college/university:	Stanford
major:	Computer Science
degree:	MS
year:	2000
college/university:	Stanford University
major:	Computer Science
degree:	PHD
year:	2002
occupation:	User Interface Designer
industry:	High Tech, Computer Networking
title:	Software Engineer
career skills:	databases, mobile networking, user interfaces, massage

someone fun-loving, easy-going...

ideal match:	someone fun-loving, easy-going...
first thing you will notice about me:	smile
height:	173 centimeters
eye color:	black
hair color:	dark brown
build:	athletic
looks:	attractive
best feature:	not on the list
turn ons:	candlelight, dancing, erotica, flirting, intelligence, skinny dipping, thunderstorms
turn offs:	too much make-up
my idea of a perfect first date:	a nice dinner with followed by dancing
from my past relationships i learned:	i don't understand women
five things I cant live without:	friends, family, dancing, music, love
in my bedroom you will find:	candles, massage oil, teddy bear, hand-cuffs, honey bunny

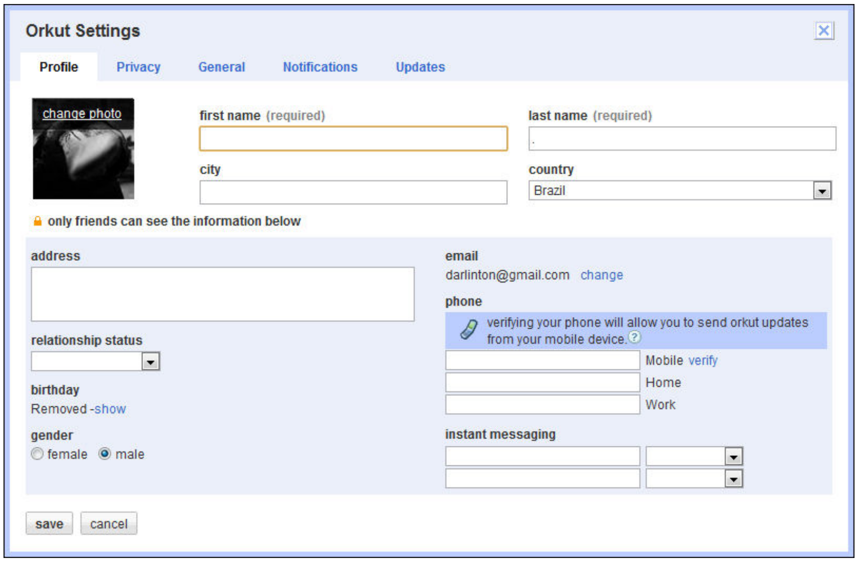
his friends (470)

Alessandra, Mario, Ricardo, Priya, Daniela, Narayan, Monique, Rodrigo, Nandini

his communities (33)

Te INCOMODO? AZAR o SEU !! (13,648); What should Google do (14,919); Clubbing in the Bay Area (2,281); Massage (52,438); Networking (90,360); Toledo Sister Cities Int'l (44)

Figure 4: Orkut User Profile Page (old version)



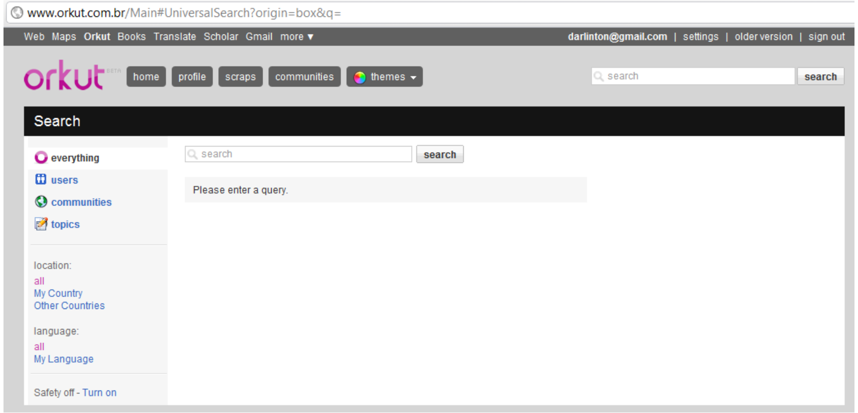
first name (required)	
last name (required)	.
city	
country	Brazil

only friends can see the information below

address	
relationship status	
birthday	Removed -show
gender	○ female ● male
email	darlinton@gmail.com change
phone	verifying your phone will allow you to send orkut updates from your mobile device
Mobile	verify
Home	
Work	
instant messaging	

Figure 5: Orkut User Profile Configuration (new version)

are two filters, one by language and another for location. The search is on the main concepts of the system and on the information that is mostly public and available for all (e.g. communities' topics). This is the main feature to search for interesting user-generated content, the social media in Orkut.



Please enter a query.

Figure 6: Orkut Universal Search

2.3 Modeling and Measurement

Based on the main concepts and features of the Orkut system, this section presents the proposed solution to exploit the Orkut social media. Regarding the research questions, this solution tries to answer the first and second questions. In this work, useful information is the public data available in Orkut. This decision has two main reasons. The first is a practical design decision for feasibility of the method implementation. The second one regards privacy issues, so the method preserves the users' privacy. Still regarding privacy, the output data from the method does not have user specific information. The focus of the analysis is on the data available in the communities.

The social media expert has to find some community of interest. This is easily achieved by the use of the Universal Search feature provided in Orkut (Figure 6). The selected community defines the domain of the analysis. The community has three main data sources: properties (e.g. description, categories), members, and forum. In this work, the additional features of the communities such as polls and events are disregarded. The proposed method makes use of the conversation in the forum and members' data as basis of the modeling and measurement of the social media.

A key design point for the measurement is the use of appealing visualization as the method outcome. The modeling and measurement of the social media generates data aggregation that is presented to the social media expert highlighting interesting characteristics in a plot.

The community forum has the users' conversations, organized by topics and posts. The model and measurement of this content uses the word cloud to summarize the content in one image. The word cloud image has the most frequent terms, with the font size related to frequency, so it is highlighting the most chatted. The word tag captures in one image the main values of the forum topics.

The members' data are used in a more sophisticated model. The user association with a community can tell a strong interest of the user in a community topic. It is important to remember that in Orkut, users employ community association as badges on their profiles, defining their preferences. Regardless the meaning of the community as a profile badge, indentifies the community relationships and generates valuable information for analysis. The Orkut system provides a feature on the community page showing up to 9 related communities (Chen, Chu, Luan, Bai, Wang & Chang 2009). This work presents a new model to unveil community relationship based on users.

The proposed model for the relationship among the communities is user-centered. The user membership on two communities establishes a relationship between these communities. So, a community is related to other if they have a user that is member of both. In a formal notation, if $user_x$ is member of cmm_a and cmm_b then cmm_a has a relationship with cmm_b. If a $user_y$ is also member of cmm_a and cmm_b then reinforce it (+1). Therefore, applying this rule over a given list of users and their membership data builds the model for the relationship among the communities. The measurement and analysis with this model are demonstrated on the case studies.

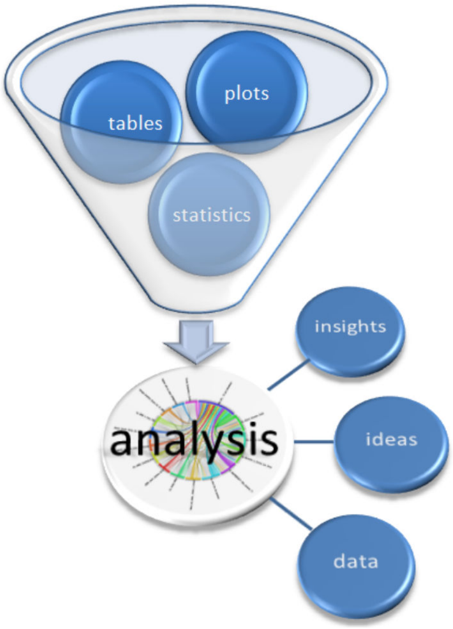


Figure 7: The social media analysis process

The outcome of this method is input for the social media expert work, or other specialist interested in the study of social media. The models and measures help the social media expert with the final analysis of the social media. Figure 7 represents the proposed process of social media analysis. Due to the kind of the data treated in this work, advanced statistics of this data can lead to a misguiding analysis, so it is not covered in this work (Krishnamurthy & Willinger 2008). An extensive discussion on the modeling and measurement of social media is presented in (Cormode et al. 2010).

2.4 The Method

The proposed method has three mains steps: data gathering, modeling/measurement, and analysis/visualization. The first step is the data gathering, which is responsible to collect the useful data from Orkut. Given the id of the community under analysis, the data gathering process retrieves the community properties (e.g. description, category), members, and forum discussion. The next step is the modeling and measurement. Following the model and measurement, as discussed in the previous section, the collected data is processed, cleaning up noise, and organized in an appropriated format for the next step. Extra data required for modeling is collected (i.e. user's communities) and processed. The analysis and visualization step takes care of showing the data in a more appealing way, with basic statistics, easing the work of the specialists executing the method.

The process starts with the gathering of the community users. However, Orkut's system limits the users displayed to only 1,000 members. A workaround for this is to use the search feature inside the community. Each search results at most 1,000 members, so making multiple searches with specific keywords can retrieve all community members. The catch in this process is to maximize the number of unique users retrieved with the minimum of searches; it is a MinMax Problem. The solution proposed is to use a list of queries (special keywords). This list of queries was built considering the alphabet, figures, common names, last names, nicknames, zodiacs, cities, states, sport teams, movies, books, and stop words.

The method implementation is split over many interactions, executed by different programs. Scripts implemented in Lua[6] process the data, while the iMacros[7] script executor on Firefox Browser[8] automates Orkut's system access. The final step of visualization plotting is done by specific programs specialized for this purpose (i.e. circos, wordle).

Wordle generates the word cloud, and Circos visualization is used for the community's relationship model, presented in the previous section. Wordle[9] is a tool for generating "word clouds" from given text. The clouds give greater prominence to words that appear more frequently in the source text. Circos is a software package for visualizing data and information. It visualizes data in a circular layout – this makes Circos ideal for exploring relationships between objects or positions (Krzywinsk, Schein, Birol, Connors, Gascoyne, Horsman, Jones & Marra 2009).

The communities relationship model building requires many steps. Due to its complexity, it is worth to detail the method execution, explaining some design choices of the method development. Bellow is the detailed steps of the communities' model.

Step 1) Generate initial user collect script
Input: (csv) search key-word (queries) file, community id (c)
Output: (iim) macro files to collect the users volume retrieved by the searches

The searches results are up to 1,000 members, but the system splits the result on pages of only up to 10 members each. In order to collect the users retrieved by the searches, the script needs to know the exact amount of users to iterate over the pages.

Step 2) Execute the macro files on iMacros to retrieve data from Orkut System

[6] Progamming language available at `http://www.lua.org/`

[7] iMacros was designed to automate repetitious tasks on the web and is available at `http://imacros.net`

[8] Internet Browser available at `http://www.mozilla.com/en-US/firefox/firefox.html`

[9] `http://www.wordle.net/`

Step 3) Generate the search scripts
Input: (`html`) collected list of users of c grouped by queries (search key-word)
Output: (`iim`) macro files to collect users of the community c

Step 4) Execute the macro files

Step 5) Extract users from the searches
Input: (`html`) query results with a list of users grouped by key-words
Output: (`db`) creation of the users' database

Step 6) Generate the user profile collect script
Input: (`db`) database of users
Output: (`iim`) macro files to collect users' profile

Step 7) Execute the macro files

Step 8) Extract users' profile
Input: (`html`) users' profile - the community association counter specially
Output: (`db`) update database of users with the communities association counter

Step 9) Generate the users' community collect script
Input: (`db`) database of users with the communities association counter
Output: (`iim`) macro files to collect users' communities

Step 10) Execute the macro files

Step 11) Extract users' communities
Input: (`html`) users' communities list
Output: (`db`) update database of users with the community info

Step 12) Create model of communities' relationship
Input: (`db`) database of users with the communities
Output: (`db`) communities' relationship model

The final step is to plot the visualization. As said earlier, the visualization is built using the Circos System. Circos has a special component to generate this kind of visualization, the Table Viewer[10]. The Figure 8 has the detailed execution of this process. The script of step 13 generates the table.txt, which is provided as input for the table viewer. It is noteworthy that for better visualization the configuration files of circos is customized for this work.

Step 13) Plot Visualization
Input: (`db`) communities' relationship model (table.txt)
Output: (`img`) visualization file

[10] `http://mkweb.bcgsc.ca/circos/presentations/articles/vis_tables2/`

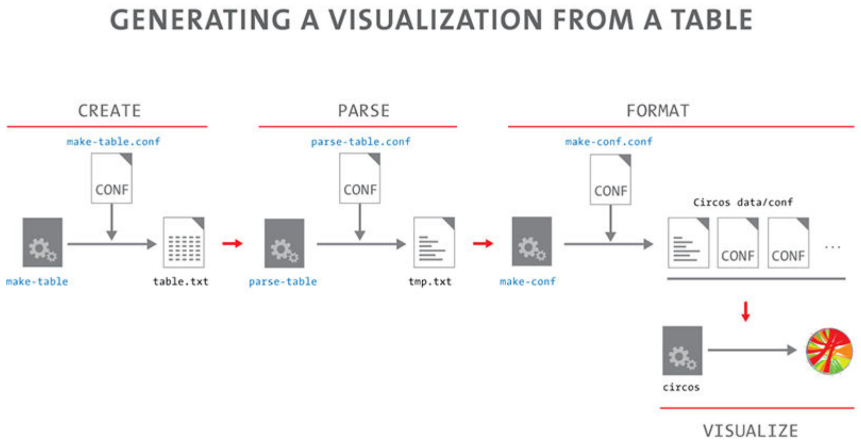


Figure 8: Table viewer process

3 Case Study

There are three case studies. The first case study is the pilot, which served as a way to outline the problem, since this work started as an exploratory work. The method design and implementation occurred along this case study execution. The focus of the first study was the Brazilian Election, a hot topic at the time of the study execution. This study showed the main issues and limitations of the method. The other two case studies focus on branding analysis, the original purpose for the method use. These studies answer the last two research questions as well.

3.1 Brazilian Elections

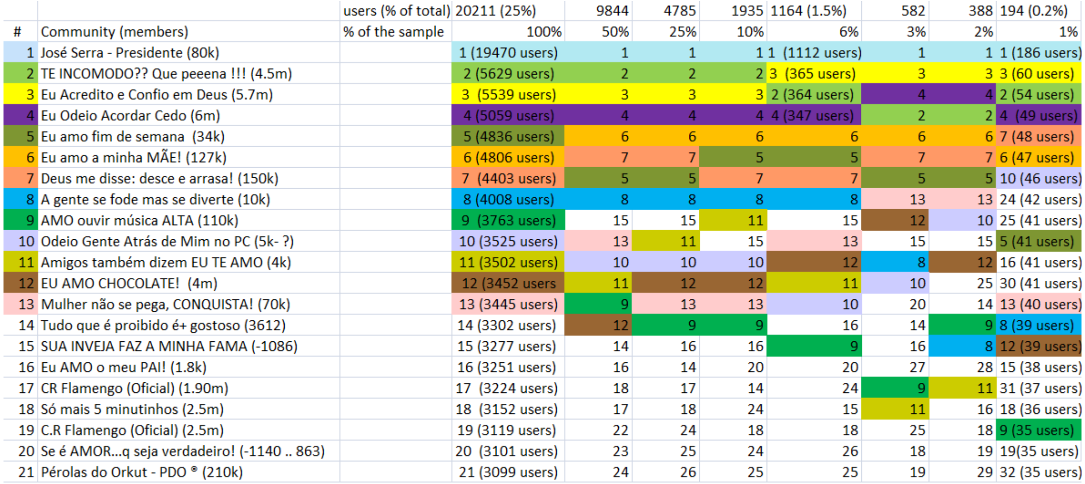
The first case study is about a general topic of interest, the Brazilian Election. The focus was on a medium-sized community (less than a 100 thousand users). The chosen community[11] was the one of the second-place candidate in the polls (José Serra), with around 80 thousand users at the time of the study (August 2010).

The scripts for the method execution were built along the case study, as an on the fly development. The results of this case study are not the best that can be produced by the method, but it shows interesting properties. The main result of this case study is to show the main issues of the method, like the processing limits, unreliable Orkut data and sampling size. The first big problem was to cope with the access to Orkut's system. The system pages construction are realized mainly by AJAX methods, so the scraping scripts must support full Javascript execution. Orkut also has a protection that limits the number of requests per second by IP number. The solution to this is to use a macro execution plug-in iMacros for Firefox. iMacros also has also some performance issues regarding the size of the macros and the generated files (HTML). After intensive experimentation, a good size ration for the files is around a thousand lines for the macro scripts and at most 40 MB for the generated file (the number of lines does not matter much).

The second big problem regards the data space of analysis. Even considering a mid-sized community, the first problem (system access by macros) plus connection problems

[11] http://www.orkut.com.br/Main#Community?cmm=355236

and Orkut System eventual unavailability troubled the study execution, leading to an incomplete data acquisition of the community members. After a month of execution, 50 thousand of 80 thousand (65%) of the users were identified, but only data of 20 thousand were retrieved. This situation also motivated a sampling study. 20 thousand users are 20% of the community users, but as shown in the data in Figure 9, the method results converge to a result with an increase of the amount of data. The analysis aims to identify related communities that describe user quality, so the order of the communities in the list is less important than the fact the community appearance in the list. Table 1 has the full list with all the communities identified in the model, and the ids are the same as those presented in Figure 9. The number between parentheses is the number of members (discussed further in the third problem). In other words, the community relationship discovery is more important than the order of importance, because the quality value is preferable to the quantity measured in the model.



		users (% of total)	20211 (25%)	9844	4785	1935	1164 (1.5%)	582	388	194 (0.2%)
#	Community (members)	% of the sample	100%	50%	25%	10%	6%	3%	2%	1%
1	José Serra - Presidente (80k)		1 (19470 users)	1	1	1	1 (1112 users)	1	1	1 (186 users)
2	TE INCOMODO?? Que peeena !!! (4.5m)		2 (5629 users)	2	2	2	3 (365 users)	3	3	3 (60 users)
3	Eu Acredito e Confio em Deus (5.7m)		3 (5539 users)	3	3	3	2 (364 users)	4	4	2 (54 users)
4	Eu Odeio Acordar Cedo (6m)		4 (5059 users)	4	4	4	4 (347 users)	2	2	4 (49 users)
5	Eu amo fim de semana (34k)		5 (4836 users)	6	6	6	6	6	6	7 (48 users)
6	Eu amo a minha MÃE! (127k)		6 (4806 users)	7	7	5	5	7	7	6 (47 users)
7	Deus me disse: desce e arrasa! (150k)		7 (4403 users)	5	5	7	7	5	5	10 (46 users)
8	A gente se fode mas se diverte (10k)		8 (4008 users)	8	8	8	8	13	13	24 (42 users)
9	AMO ouvir música ALTA (110k)		9 (3763 users)	15	15	11	15	12	10	25 (41 users)
10	Odeio Gente Atrás de Mim no PC (5k- ?)		10 (3525 users)	13	11	15	13	15	15	5 (41 users)
11	Amigos também dizem EU TE AMO (4k)		11 (3502 users)	10	10	10	12	8	12	16 (41 users)
12	EU AMO CHOCOLATE! (4m)		12 (3452 users	11	12	12	11	10	25	30 (41 users)
13	Mulher não se pega, CONQUISTA! (70k)		13 (3445 users)	9	13	13	10	20	14	13 (40 users)
14	Tudo que é proibido é+ gostoso (3612)		14 (3302 users)	12	9	9	16	14	9	8 (39 users)
15	SUA INVEJA FAZ A MINHA FAMA (-1086)		15 (3277 users)	14	16	16	9	16	8	12 (39 users)
16	Eu AMO o meu PAI! (1.8k)		16 (3251 users)	16	14	20	20	27	28	15 (38 users)
17	CR Flamengo (Oficial) (1.90m)		17 (3224 users)	18	17	14	24	9	11	31 (37 users)
18	Só mais 5 minutinhos (2.5m)		18 (3152 users)	17	18	24	15	11	16	18 (36 users)
19	C.R Flamengo (Oficial) (2.5m)		19 (3119 users)	22	24	18	18	25	18	9 (35 users)
20	Se é AMOR...q seja verdadeiro! (-1140 .. 863)		20 (3101 users)	23	25	24	26	18	19	19(35 users)
21	Pérolas do Orkut - PDO ® (210k)		21 (3099 users)	24	26	25	25	19	29	32 (35 users)

Figure 9: Sampling study

The sampling study is an answer to the third research questions and shows that even with very small samples the model can provide a good hint of the result of the full sample analysis. This sampling study also shows a convergence of the modeling result with and increase of the analyzed data. Therefore, the method can deliver reasonably good analysis even in the early stages of execution and the results improve with the data scraping during the process.

The third problem still is an open problem, because it is problem with Orkut's system. The system provides unstable and unreliable counters. An important counter for the method is the number of community members. However, Orkut's system has a problem with this counter, returning wrong amounts of members (sometimes returning negative values). Since it is a problem with Orkut, it is hopeless to try a solution by the method. This is the main motivation to invalidate more sophisticated statistics analysis of the collected data. It is very hard to determine the full amount of the available data, and to retrieve it all.

At the end, this case study analysis considered data of 20,211 users, with 1,055,867 communities in the model. The communities' relationships file in plain text with 5,879,580 pairs of community id and its relationship weights has up to 17 GB. Figure 10 has the

1	José Serra - Presidente (80k)
2	TE INCOMODO?? Que peeena !!! (4.5m)
3	Eu Acredito e Confio em Deus (5.7m)
4	Eu Odeio Acordar Cedo (6m)
5	Eu amo fim de semana (34k)
6	Eu amo a minha MÃE! (127k)
7	Deus me disse: desce e arrasa! (150k)
8	A gente se fode mas se diverte (10k)
9	AMO ouvir música ALTA (110k)
10	Odeio Gente Atrás de Mim no PC (5k- ?)
11	Amigos também dizem EU TE AMO (4k)
12	EU AMO CHOCOLATE! (4m)
13	Mulher não se pega, CONQUISTA! (70k)
14	Tudo que é proibido é + gostoso (3612)
15	SUA INVEJA FAZ A MINHA FAMA (-1086)
16	Eu AMO o meu PAI! (1.8k)
17	CR Flamengo (Oficial) (1.90m)
18	Só mais 5 minutinhos (2.5m)
19	C.R Flamengo (Oficial) (2.5m)
20	Se é AMOR...q seja verdadeiro! (-1140 .. 863)
21	Pérolas do Orkut - PDO (210k)
22	A Preguiça mata? Morri!
23	Eu Tenho MSN !
24	TE PERGUNTEI alguma Coisa?
25	Já tentei voltar p/mesmo sonho
26	Tocava Campainha e Corria
27	Nada Acontece Por Acaso
28	Se é AMOR...q seja verdadeiro!
29	Tem dias que tudo me irrita!
30	Dou RISADA Quando Não Pode!
31	Putz, para de encher o saco
32	Ouço mil VEZES a mesma MÚSICA!

Table 1: Extended list of the top communities

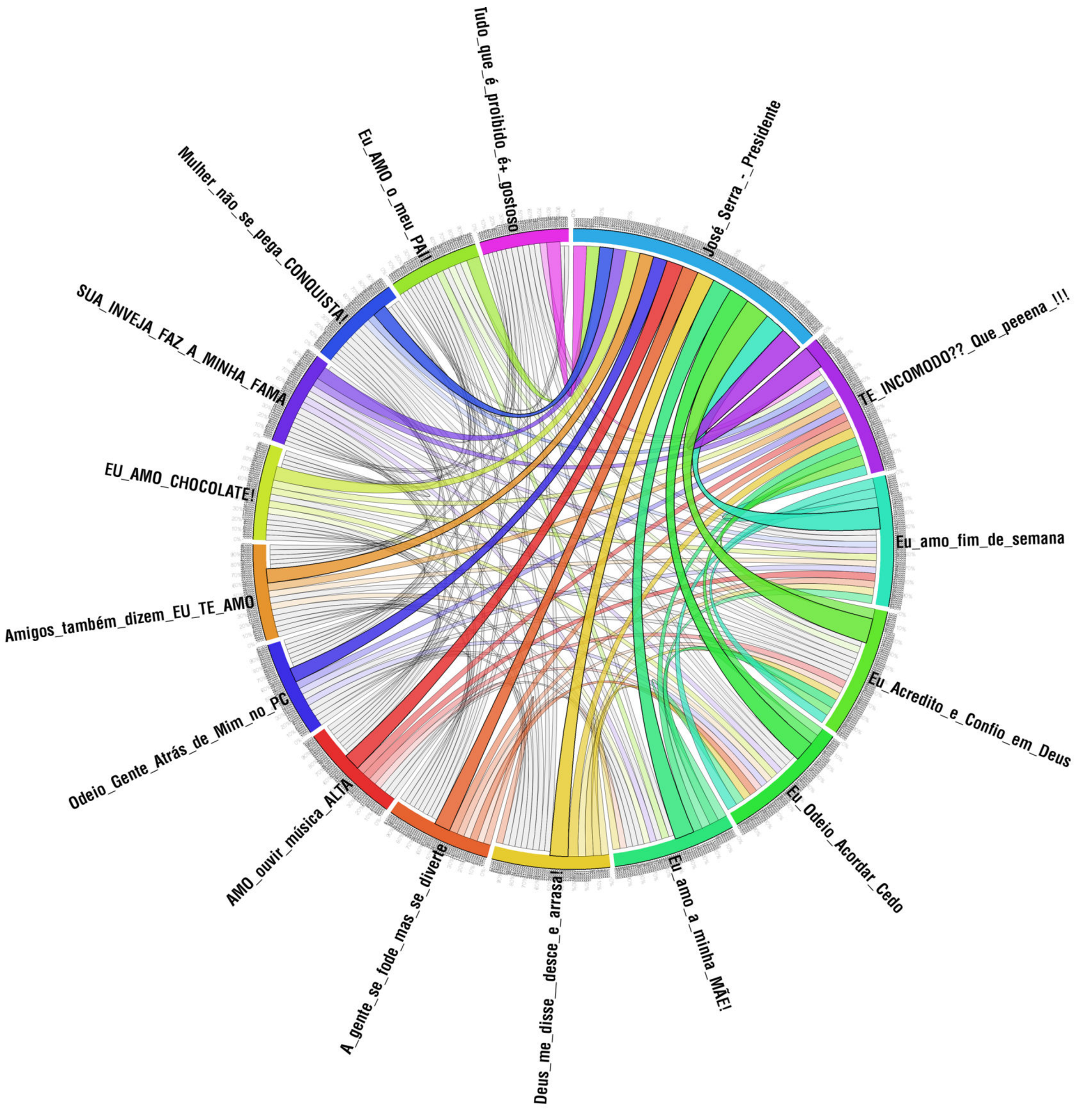


Figure 10: Visualization of the top 16 communities of the model

final plot of the analysis. The generation of the visualization is achieved using Circos, as previously described in Section 2.4. A possible use of the graph by the social media specialist is to identify user values as religion, family, food taste, music preferences, self-esteem and confidence. The next case studies show a better way to scrutinize these values, considering it with specific categories of communities instead of this mixed model.

3.2 Brand Analysis: Sky

The second case study is a brand analysis of Sky, the Brazilian leader of Paid TV. The main community about the company in Orkut is the accessible by the URL `http://www.orkut.com.br/Main#Community?cmm=5233175`. It has around 30 thousand users (October 2010).

The method execution begins by retrieving the community users. The term files used for the members search has 17,920 entries. From these terms, 7,488 returned zero user results. From the other 10,432 successful searches, 19,267 (63% of total) were identified. The model of communities for these users has 1,073,784 communities with 5,079,972 relationships. The full method execution took a couple of weeks, and the analysis considers all the reachable data.

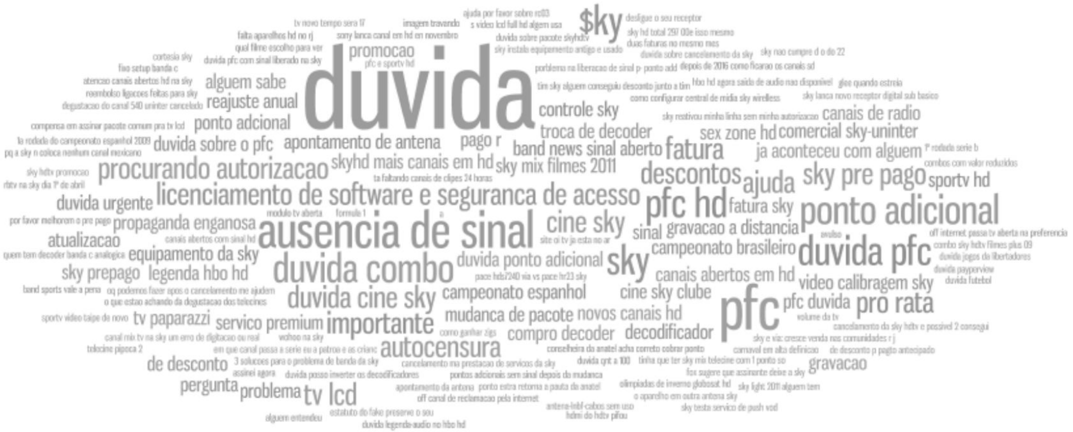


Figure 11: Tag cloud of the forum posts content

The word clouds of the forum topics are shown in Figure 11. The main word used in the topics is "question" ("duvida" in Portuguese), followed by "signal problems" ("ausencia de sinal") and the soccer channel "pfc". This picture can quickly provide the main use of the community; it is a forum for helping the company's clients. The model of communities tries to reveal more about the related communities and the user's preferences. Figure 12 shows the plot for the model considering all communities. From social media expert feedback, the method evolved to support the model building considering only a specific community category. Extra visualization plots are available in the Appendix. For example, Figure 15 has the preferable TV programs and shows; the Figure 18 has the most important Brazilian soccer teams (Flamengo, São Paulo and a minor appearance of Corinthians) and the sport-brand Adidas; Figure 16 shows the user identification of Ayrton Senna and Felipe Massa, but no mention of Rubens Barrichello. The detailed list of the communities model are presented in Table 2, community id being the first column, member counter being the second column, and the community name with the number of users associations identified in the model between parentheses in the third column.

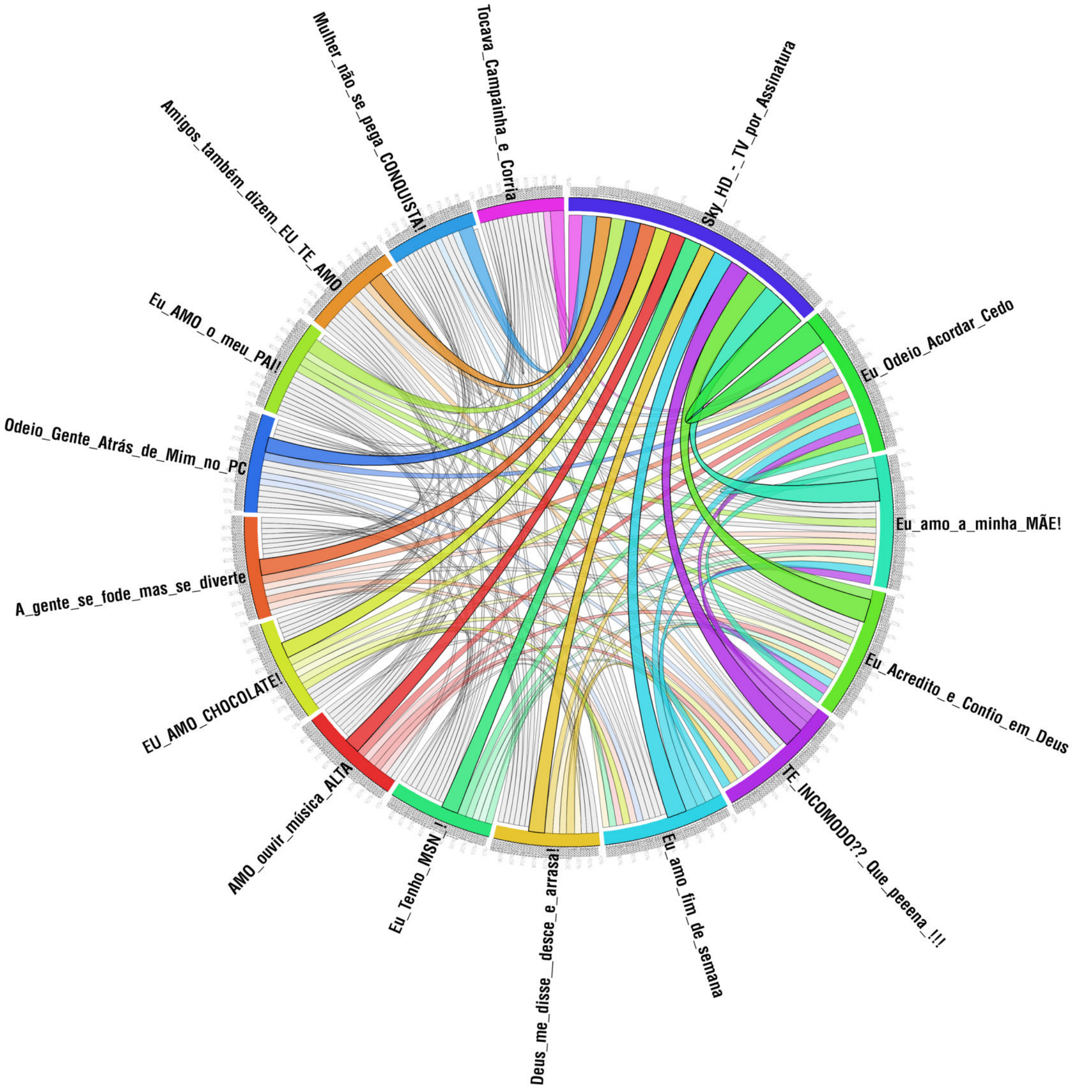


Figure 12: Model of communities' relationship of the Sky Case Study

This case study is a good example of the method outcome and gives a good hint of the answer for the fourth and last research question. It is clear that the method needs to improve much more to provide an extensive answer to the proposed research question, but so far, the analysis provided by the method tends to extract qualitative attributes from the available social media.

3.3 Brand Analysis: BlackBerry

The third case study is a brand analysis of Blackberry, a global big player in smartphones. The main community about the company in Orkut is accessible by the URL `http://www.orkut.com.br/Main#Community?cmm=13637019`. It is interesting to point out that the

id	members	community (#users)
5233175	30488	Sky HD - TV por Assinatura (19057 users)
68685	6163537	Eu Odeio Acordar Cedo (4692 users)
342550	5884789	Eu Acredito e Confio em Deus (4183 users)
176183	139649	Eu amo a minha MÃE! (4058 users)
1771742	4663443	TE INCOMODO?? Que peeena !!! (3513 users)
823066	3567441	Eu amo fim de semana (3497 users)
68801	-3224	Eu Tenho MSN ! (3033 users)
61349	16136	A gente se fode mas se diverte (2878 users)
926	150089	AMO ouvir música ALTA (2853 users)
10320499	196925	Deus me disse: desce e arrasa! (2849 users)
42078	3993847	EU AMO CHOCOLATE! (2821 users)
748810	9836	Odeio Gente Atrás de Mim no PC (2710 users)
288442	598	Amigos também dizem EU TE AMO (2671 users)
226394	11875	Eu AMO o meu PAI! (2611 users)
4581547	74463	Mulher não se pega, CONQUISTA! (2582 users)
264826	1577072	Tocava Campainha e Corria (2462 users)

Table 2: Extended list of the top communities from the model for the Sky Case Study

main community by far is Brazilian, even though it is a Canadian company with many more users outside Brazil. It has around 8 thousand users (November 2010).

The method was able to retrieve 7,194 users (90% of total). The model of communities for these users has 481,855 communities with 1,551,173 relationships. The full method execution took a couple of weeks, and the analysis considers all the reachable data.

The word clouds of the forum topics are shown in Figure 13 . This picture can quickly provide the main use of the community; it is a forum of users looking for help. There is also some topics on specific phone models and commerce (selling). Figure 14 shows the plot for the model considering all communities. As pointed out in the previous case study by the social media experts, the model with all communities helps only a little in the analysis.

Extra visualization plots are available in the Appendix. For example, Figure 19 has the main company communities related to Blackberry's community, and Figure 22 shows some fashion companies relationships as Armani with Dolce & Gabana and Volcom with Billabong. Figure 20 displays only communities related with computers and Internet, with popular blackberry models and iPhone from Apple (the main competitor). Figure 21 tells about user's location, having the unexpected result of Rio de Janeiro with more relationships than São Paulo (the biggest city and with more businesses).

For a more elaborate analysis, the social media experts asked for a tool to analyze the users' posts, to review the conversation in depth. It is desirable to have this tool with a rich graphic visualization, given the big picture in a glance, but also enabling them to go beyond of the pure text analysis. So far, the method is good to reveal major trends of the social media available in Orkut, but further investigation is needed to enable the specialists to scrutinize the data deeply.

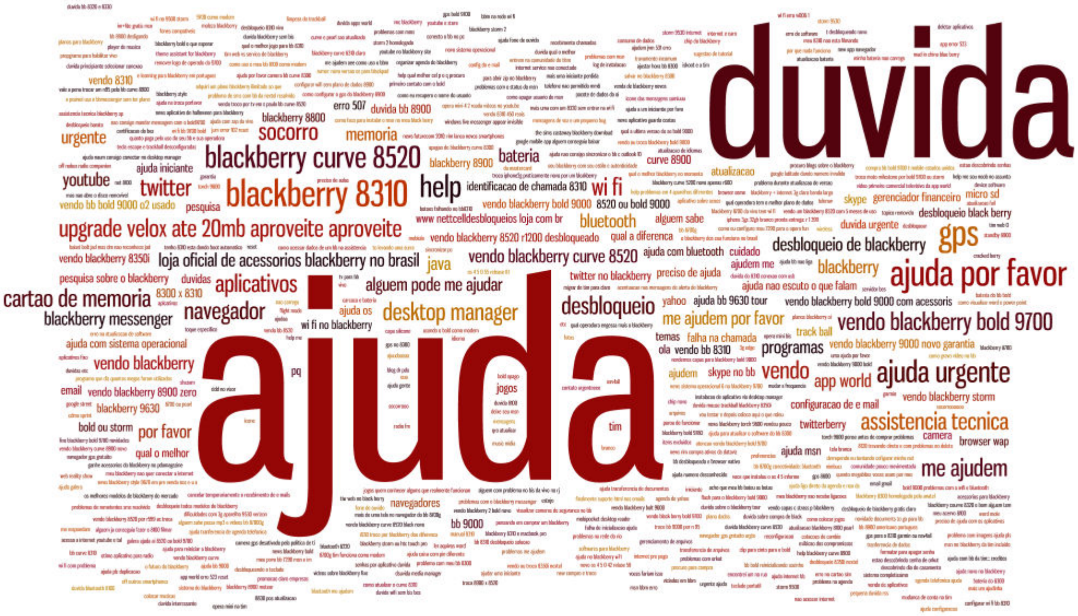


Figure 13: Word cloud of the forum topics of the Blackberry community

4 Conclusion

This work presents a new method to exploit the social media available in Orkut, the leading "social network" website in Brazil. Exploratory research questions guide the work development. Besides the research questions, there are several key-features desirable for the proposed method, as privacy concerns, domain-oriented modeling, quality than quantity measurements, and rich-graphic data visualization.

The method design aims to make use at the public data of Orkut, composed mainly of communities' content, but also respecting the user privacy. The epistemology considered in the development of the method is discussed, being the method foundation domain-oriented (i.e. Orkut main concepts) and the others key-features of design. The model and measurements used on the method are defined as well. The method use is demonstrated by three cases.

According to (Cormode et al. 2010), the method is qualified as follows. It is a scraping based for data collection, due to Orkut's system limitations. The sampling methodology is an exhaustive crawling within a defined boundary, having a starting point in a community of interest. It is also shown that the method works with small samples, and it converges to better representations if more data are considered in the measurement. The measurement efforts focus on quality attributes from data rather than statistics.

The main contribution of this work is the new proposed method. Although, another important outcome from this research is to ground future work's development. There are two main paths for this work's further development. The first is to improve the data analysis with more methods to analyze the users' discussion on the communities. Another way is to support new social media sources such as Facebook, Twitter, and the blogosphere in general, creating a platform to support specialists in the social media analysis in general.

The ongoing research project focuses on building a multi-agent system as a solution for

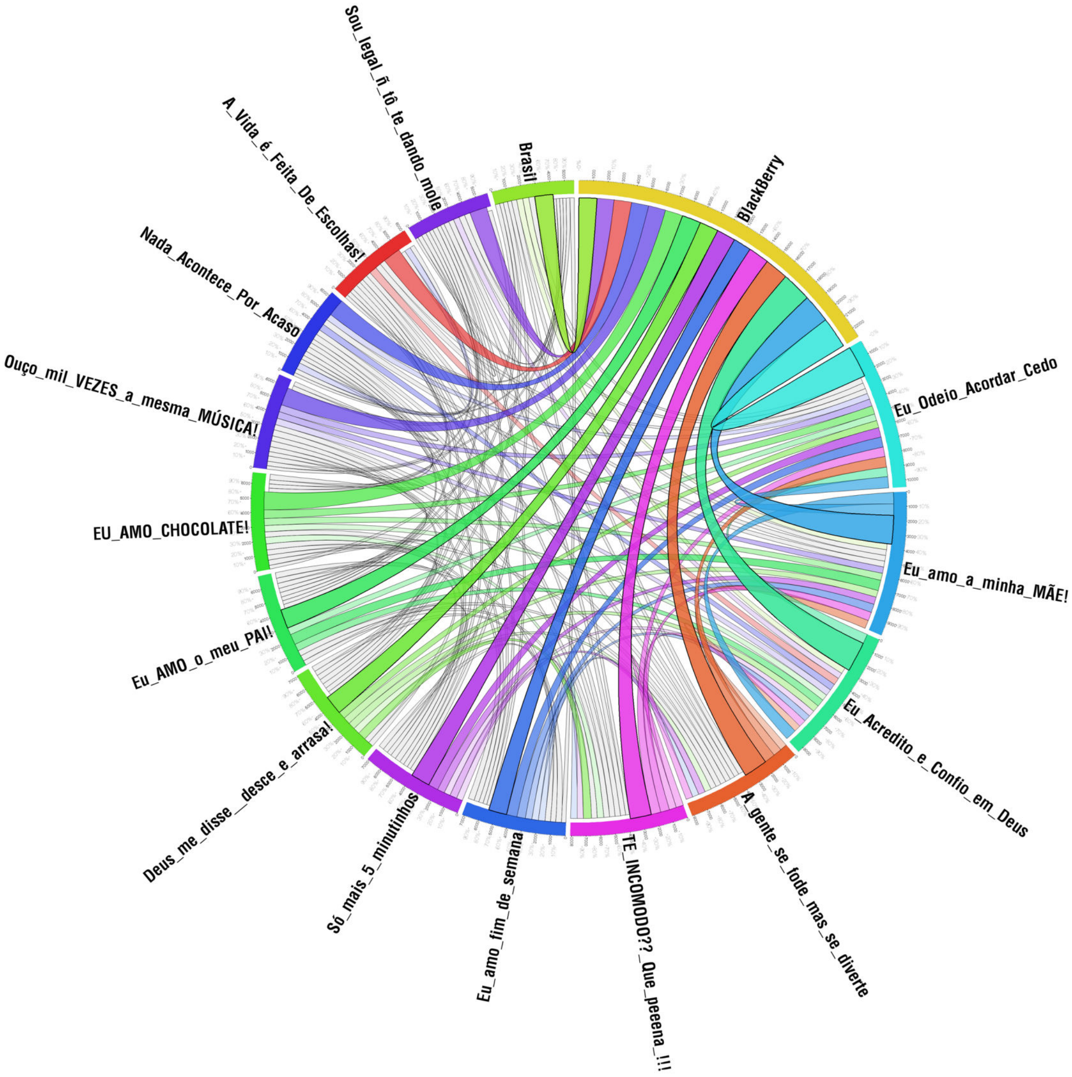


Figure 14: Model of communities' relationship of the Blackberry Case Study

the problem to study the social data. The main design goals are as follows. I) Governance: empower the users as the “king” of the system. II) Architecture: composed by flexible, smart and adapted components. III) Strategic sourcing: find and use strategic sources of social media data. IV) Processes automation: automation of manual tasks as much as possible. V) Meaningful visualization: the process outcome must be easily understandable, seeding insights and new ideas for analysts. The multi-agent system paradigm addresses the main design goals, for that reason the software engineering of the platform to research social data is on a multi-agent system.

5 Acknowledgments

The authors would like to thank Rodrigo Pazzini for his help as social media expert. This work has been sponsored by the Ministério de Ciência e Tecnologia da Presidência da República Federativa do Brasil – Bolsa de Doutorado CNPq 142620/2009-2.

References

Ahmed, N. K., Berchmans, F., Neville, J. & Kompella, R. (2010), Time-based sampling of social network activity graphs, *in* ‘Proceedings of the Eighth Workshop on Mining and Learning with Graphs’, MLG ’10, ACM, New York, NY, USA, pp. 1–9.

Benevenuto, F., Rodrigues, T., Cha, M. & Almeida, V. (2009), Characterizing user behavior in online social networks, *in* ‘Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference’, IMC ’09, ACM, New York, NY, USA, pp. 49–62.

Bigonha, C. A. S., Cardoso, T. N. C., Moro, M. M., Almeida, V. A. F. & Goncalves, M. A. (2010), Detecting evangelists and detractors on twitter, *in* ‘Anais do Simpï¿½sio Brasileiro de Sistemas Multimï¿½dia e Web’, Belo Horizonte, Brazil.

Chen, W.-Y., Chu, J.-C., Luan, J., Bai, H., Wang, Y. & Chang, E. Y. (2009), Collaborative filtering for orkut communities: discovery of user latent behavior, *in* ‘Proceedings of the 18th international conference on World wide web’, WWW ’09, ACM, New York, NY, USA, pp. 681–690.

Cheng, X., Dale, C. & Liu, J. (2008), Statistics and Social Network of YouTube Videos, *in* ‘Quality of Service, 2008. IWQoS 2008. 16th International Workshop on’, pp. 229–238.

Cormode, G., Krishnamurthy, B. & Willinger, W. (2010), ‘A manifesto for modeling and measurement in social media’, *First Monday* **15**(9).

da F. Costa, L., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. (2007), ‘Characterization of complex networks: A survey of measurements’, *Advances In Physics* **56**, 167.

Huberman, B., Romero, D. & Wu, F. (2008), ‘Social networks that matter: Twitter under the microscope’, *First Monday* **14**(1).

Krishnamurthy, B. & Willinger, W. (2008), ‘What are our standards for validation of measurement-based networking research?’, *SIGMETRICS Perform. Eval. Rev.* **36**, 64–69.

Krzywinsk, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. (2009), ‘Circos: an information aesthetic for comparative genomics’, *Genome Res.* **19**(9), 1639–1645.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010), What is twitter, a social network or a news media?, *in* ‘Proceedings of the 19th international conference on World wide web’, WWW ’10, ACM, New York, NY, USA, pp. 591–600.

Nazir, A., Raza, S. & Chuah, C.-N. (2008), Unveiling facebook: a measurement study of social network based applications, *in* ‘Proceedings of the 8th ACM SIGCOMM conference on Internet measurement’, IMC ’08, ACM, New York, NY, USA, pp. 43–56.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. & Zhao, B. Y. (2009), User interactions in social networks and their implications, *in* ‘Proceedings of the 4th ACM European conference on Computer systems’, EuroSys ’09, ACM, New York, NY, USA, pp. 205–218.

A Extra visualization plots

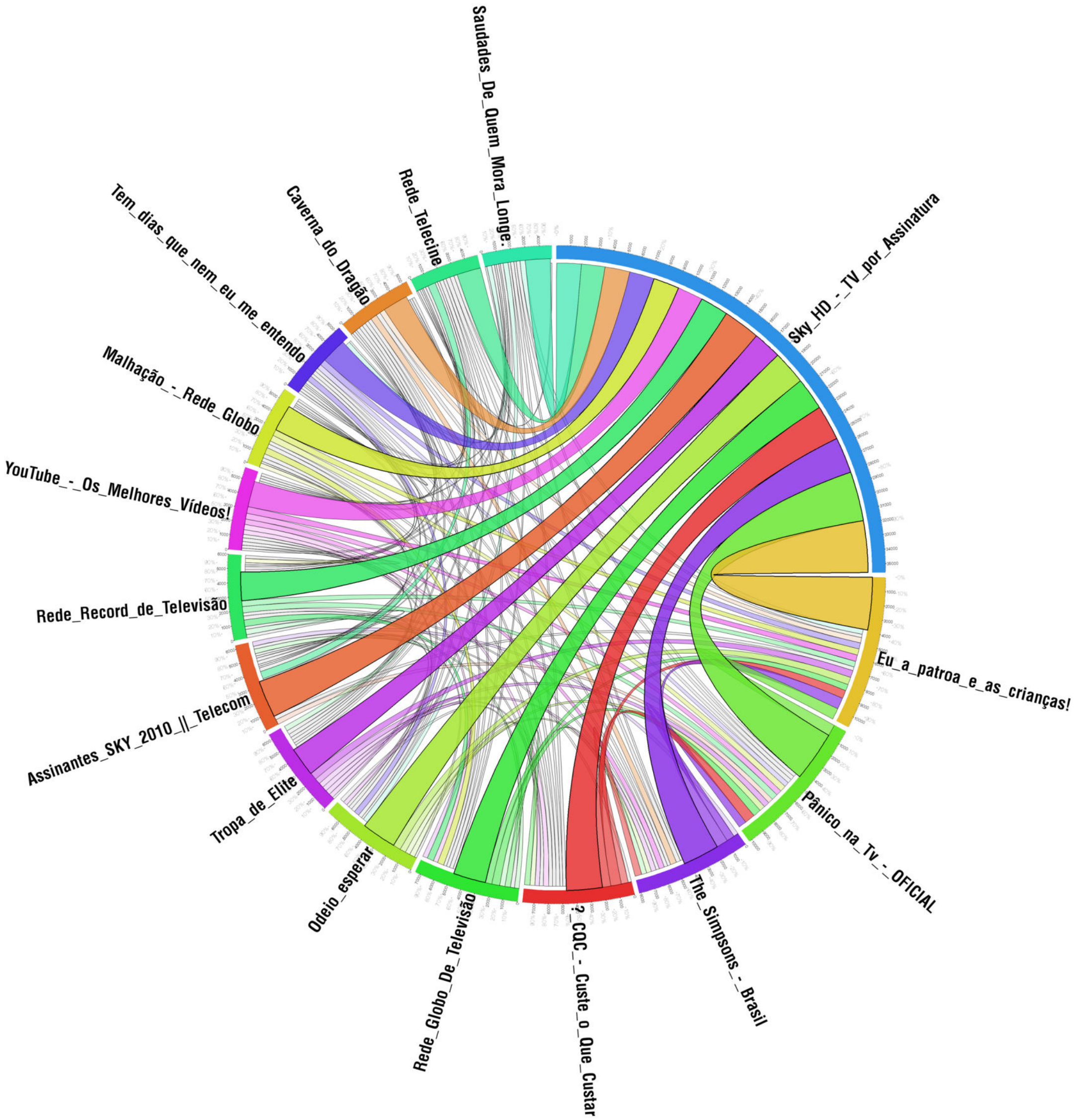


Figure 15: Arts and Entertainment Category (Sky Case Study)

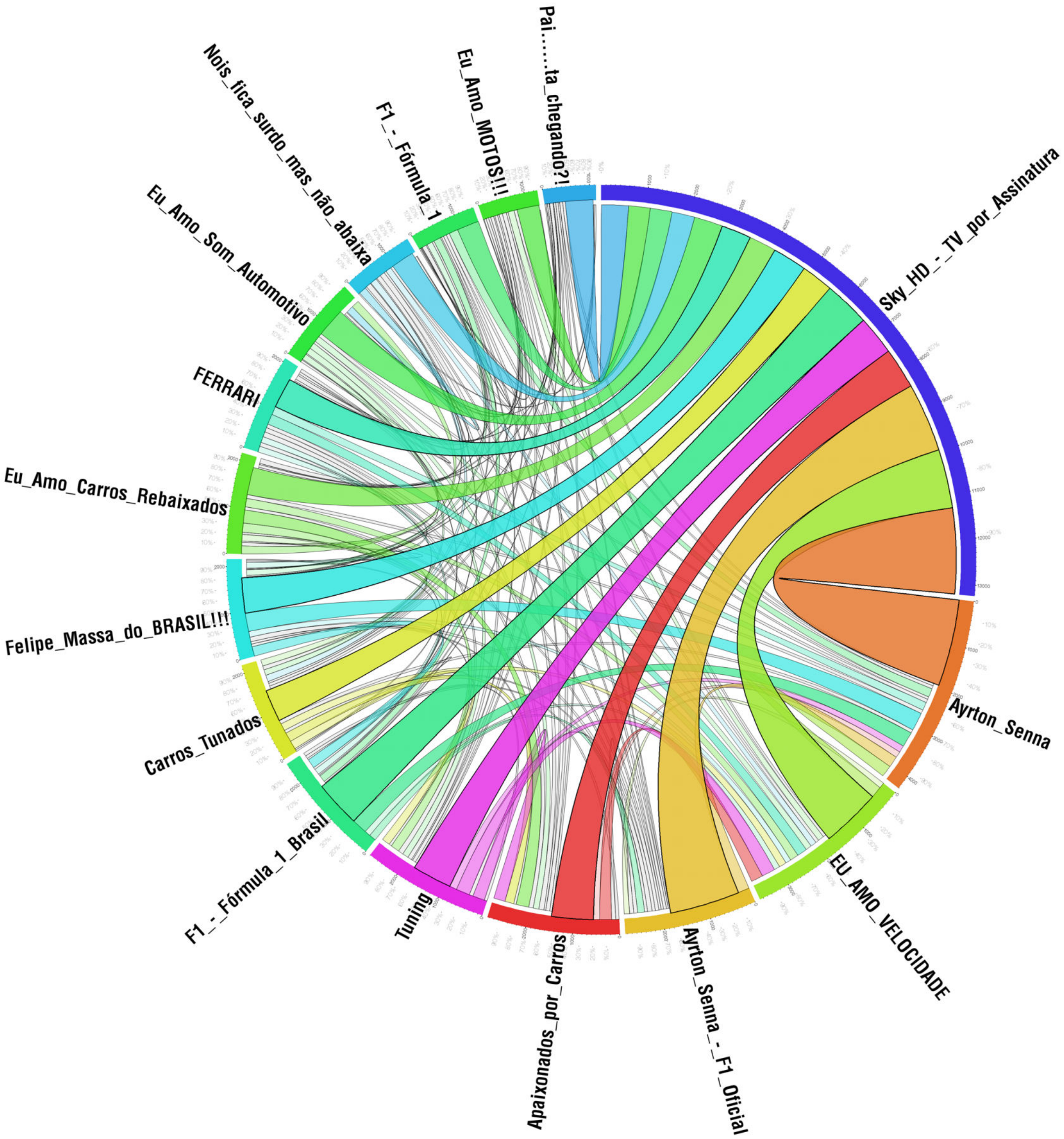


Figure 16: Automotive Category (Sky Case Study)

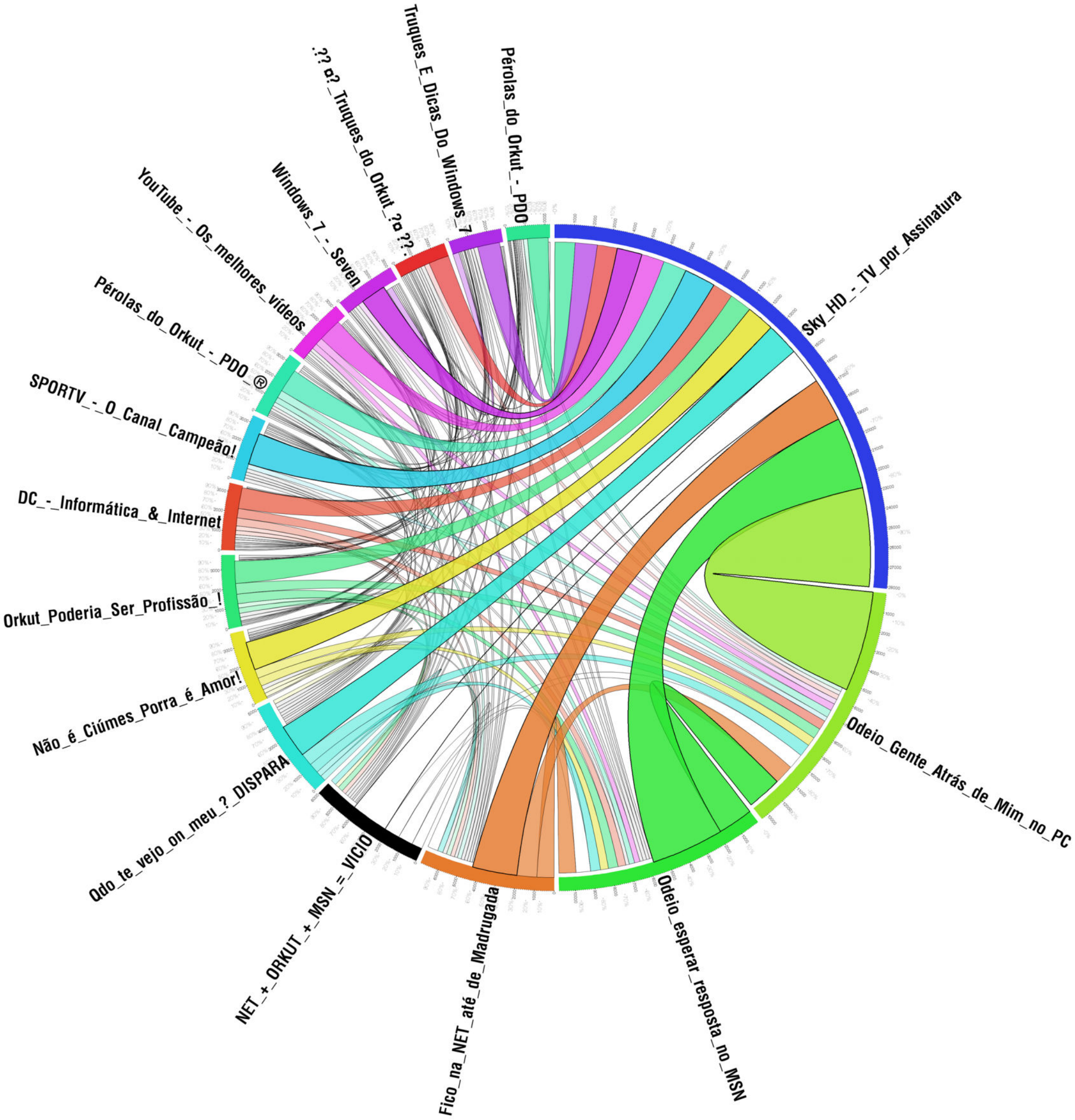


Figure 17: Internet and Computers Category (Sky Case Study)

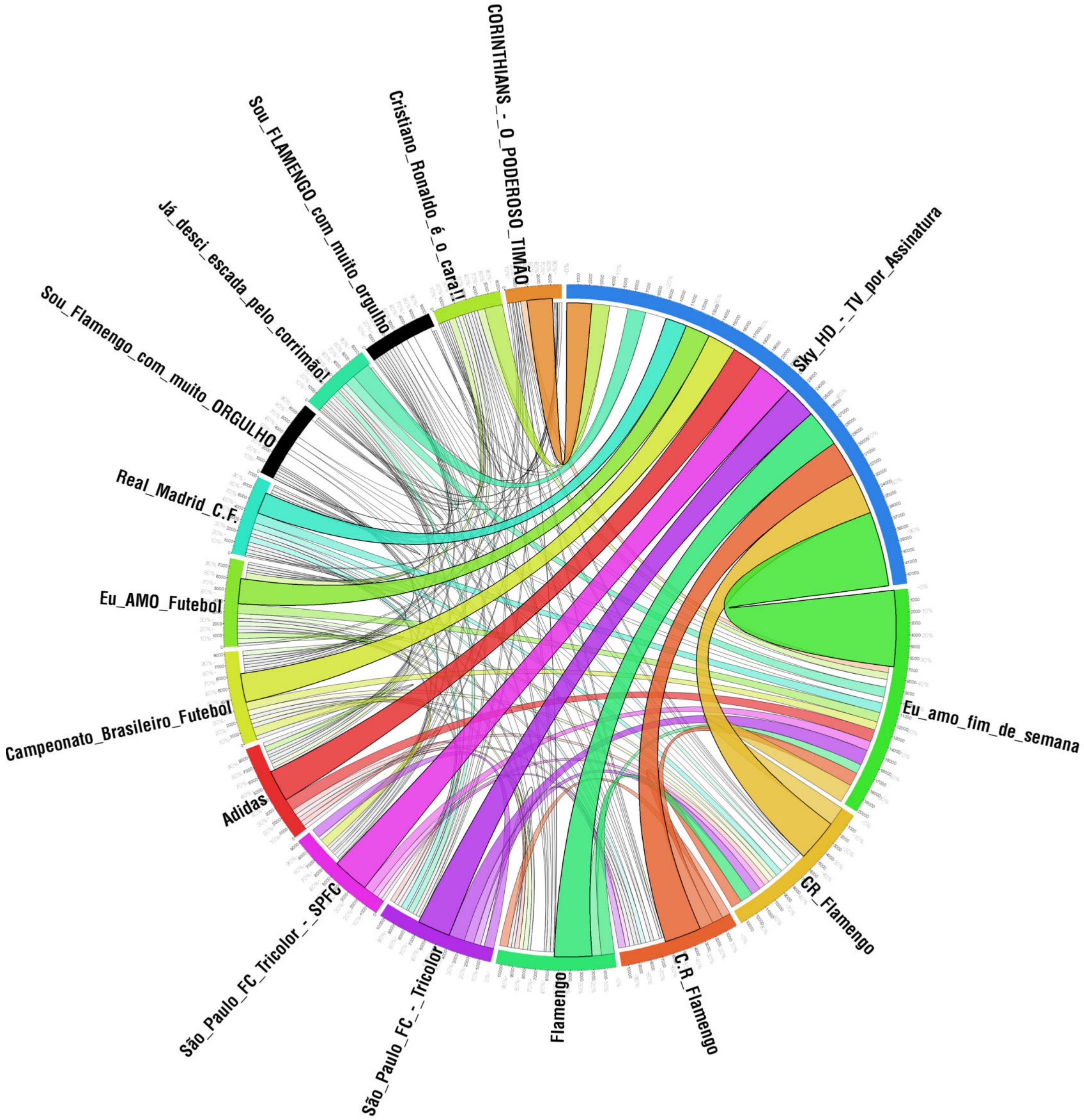


Figure 18: Sports Category (Sky Case Study)

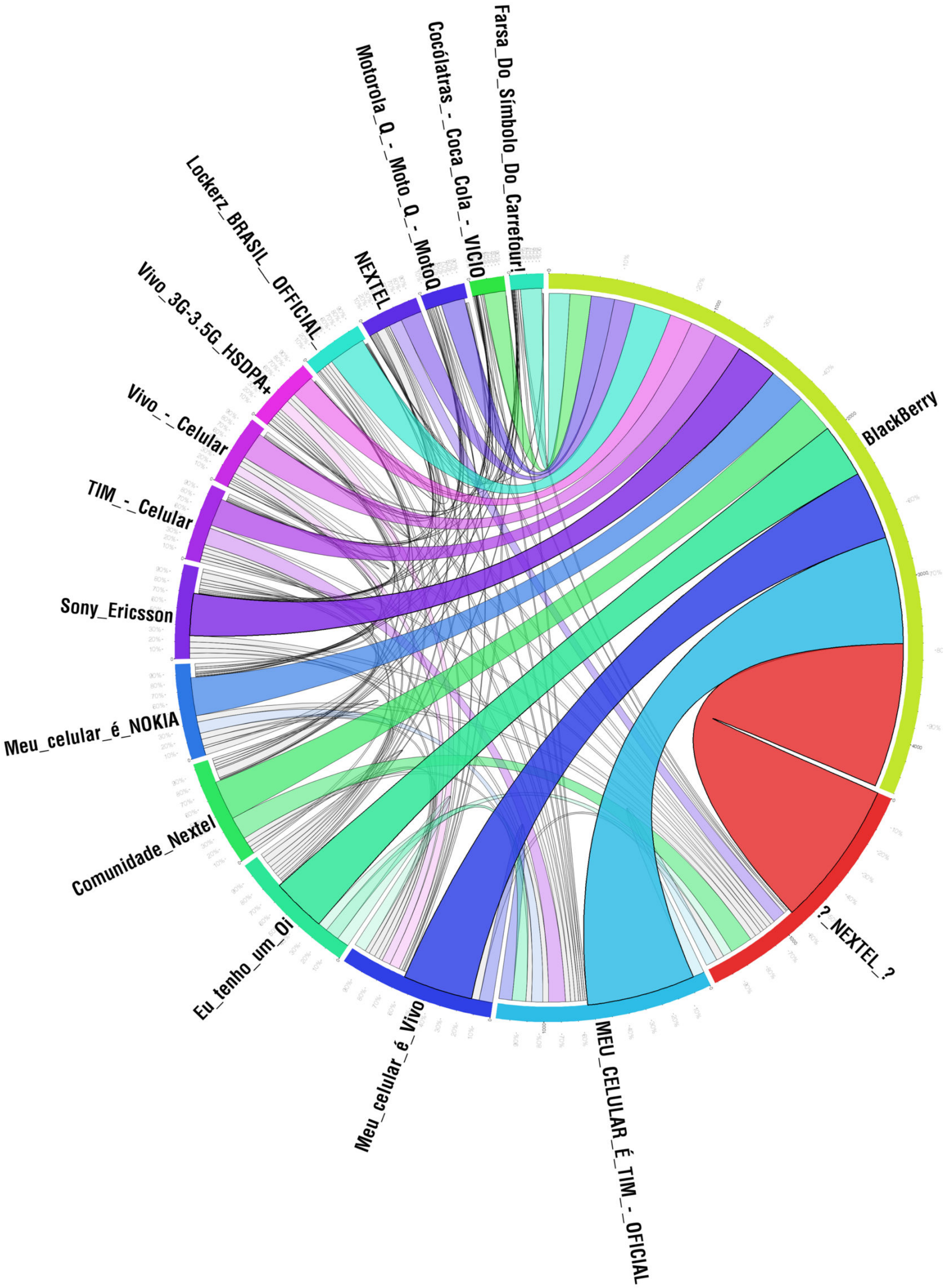


Figure 19: Company Category (Blackberry Case Study)

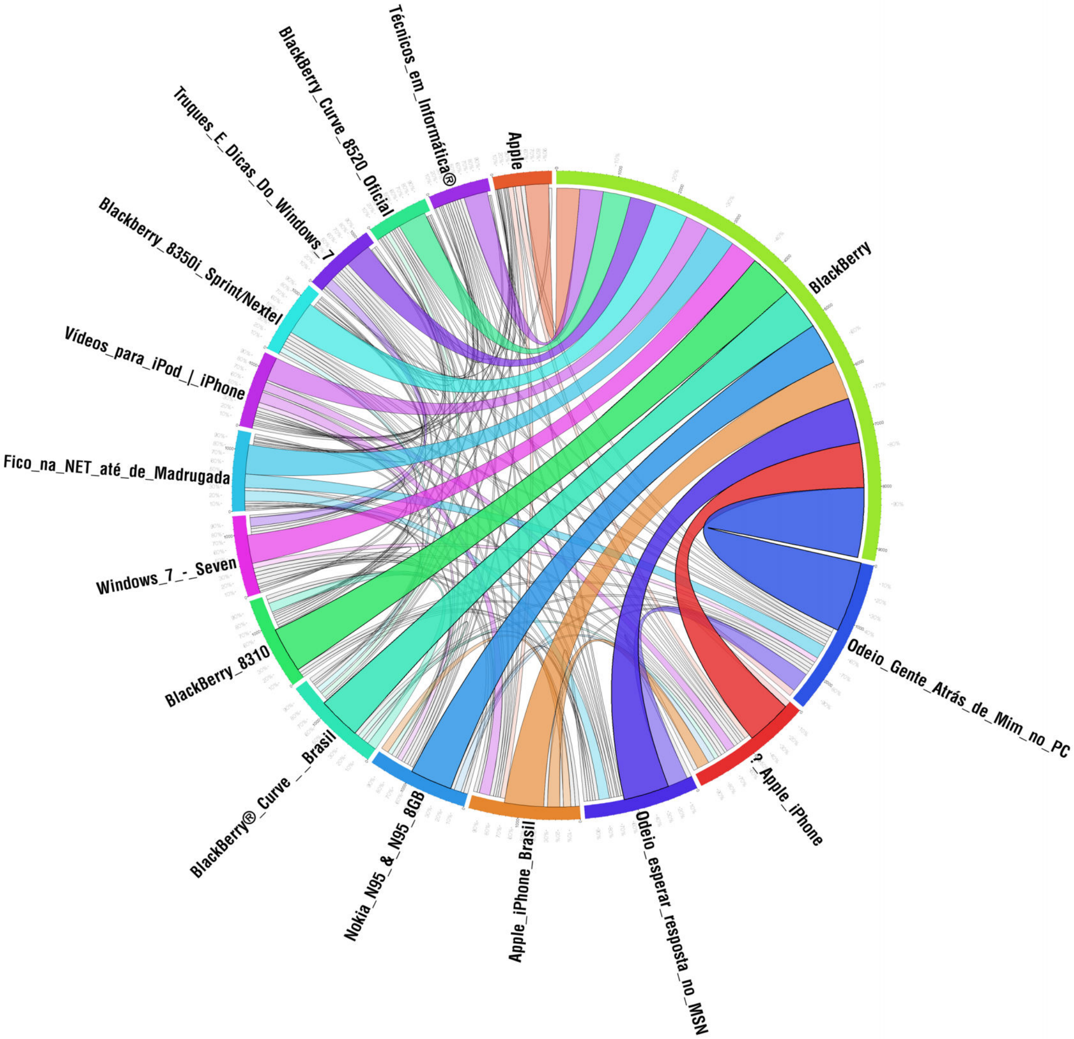


Figure 20: Internet and Computers Category (Blackberry Case Study)

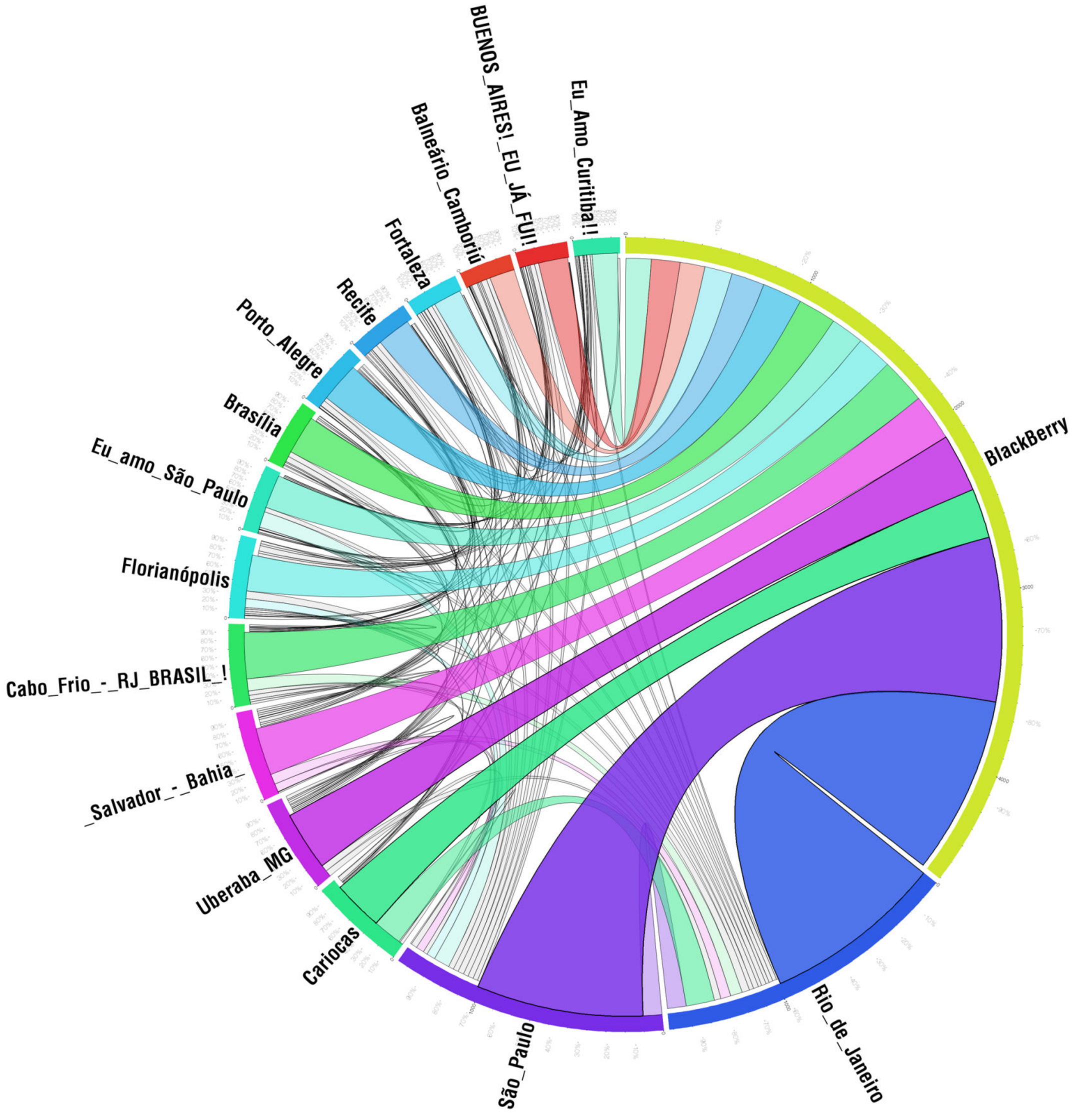


Figure 21: City and Neighborhood Category (Blackberry Case Study)

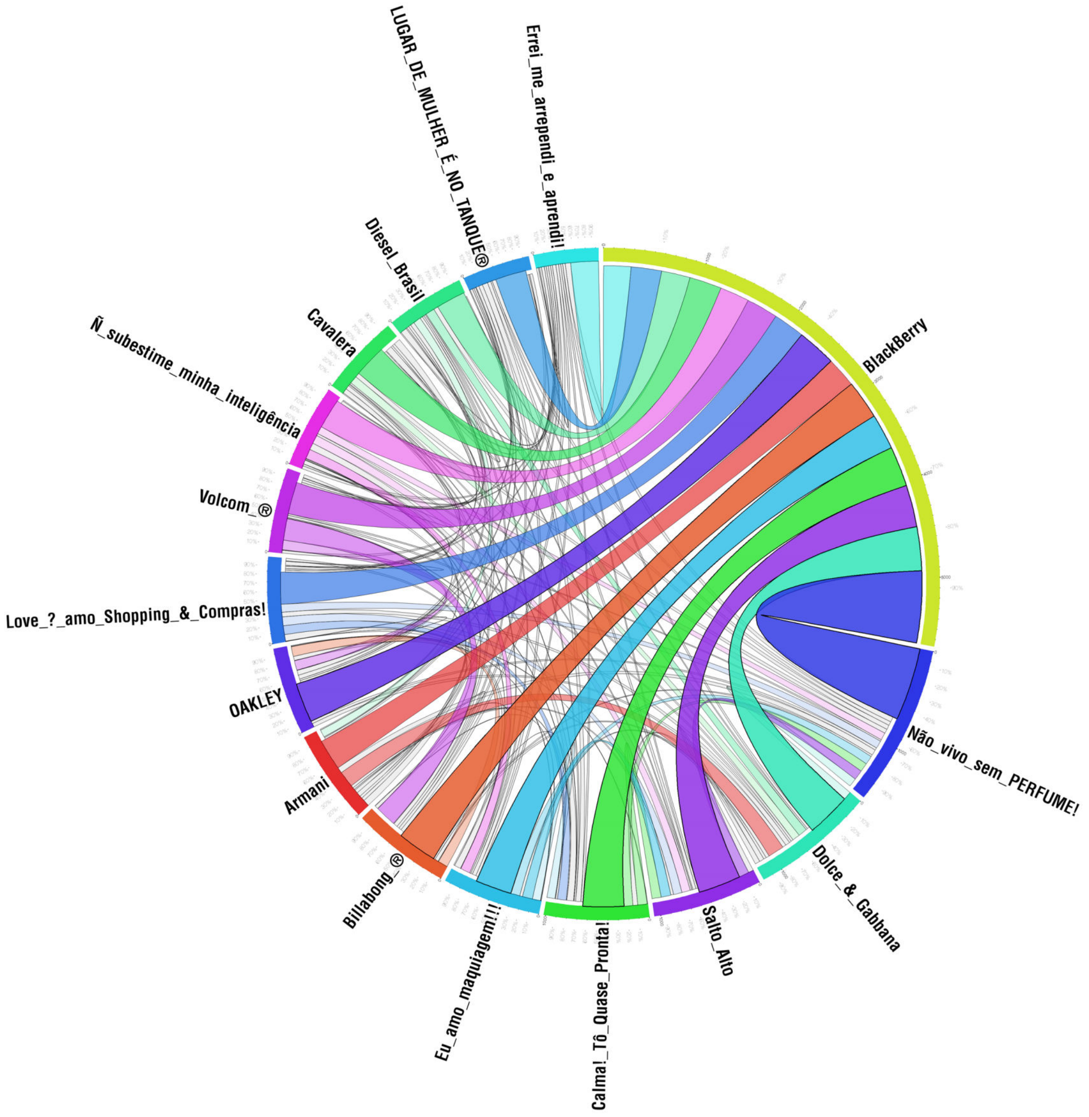


Figure 22: Fashion and beauty Category (Blackberry Case Study)