Research Digest

Telefónica Digital

Telefónica

Foreword by Carlos Domingo

During the past 20 years, the telecommunications sector has changed radically, becoming a dynamic, global and highly competitive market. In this complex world, it is essential to promote the development of differential technology that will enable our company to offer products, which will allow us to differentiate ourselves from our competitors. This is the road to create and sustain true competitive advantages.

For that reason, in 2006 scientific groups were created at Telefónica I+D, with the purpose of breaking some barriers of the traditional industrial process by attracting researchers to Telefónica. This research organisation has allowed Telefónica to be part of the knowledge network of the international scientific world, enabling the translation of scientific advances into marketable innovations.

Now within Telefónica Digital, the research organisation focuses on mid to long-term research problems related to several technological areas of interest to the Telefónica Group. It follows an open research model in collaboration with universities and other research institutions, and favours the dissemination of their work both through publications and technology transfer. Scientific publications are a good benchmark of the quality of the work produced and is a good filter to determine what is relevant and what is not. At the same time, they enable scientific knowledge evolution and encourage new advances.

With this publication, we have compiled all our 2011 research papers with the aim of further promoting its disclosure. I believe it constitutes a good reference to getting to know the state of the art of the technology in the digital world.

Carlos Domingo

Preface by Pablo Rodríguez

The mission of the Telefonica I+D Research groups is three-fold: (1) lead the early stages of the innovation funnel with expert knowledge and scientific expertise, (2) generate intellectual property rights for Telefonica, and (3) create innovative technologies and services that will enable Telefonica to maintain its leading position in the telecommunications market and be a strong player in the digital world as a true innovator. The focus is on technologies that are likely to impact Telefonica's service offerings in 3-5 years time.

The Telefónica I+D Research group was created in 2006 and follows an open research model in collaboration with universities and other research institutions, promoting the dissemination of scientific results both through publications in top-tier peer-reviewed international journals and conferences and technology transfer.

The Research organization is led by Pablo Rodriguez and includes a multi-disciplinary team. The Research leadership is shared with Nuria Oliver and Dina Papagiannaki. This multi-disciplinary and international research group comprise 20 full time researchers, holding PhD degrees in computer science, and spanning specialties from computer networking, multimedia data analysis, human computer interaction, mobile computing, distributed systems, user modeling and data mining. It also includes forty PhD interns and visitors from universities across the world each year.

Within the past 5 years, the achievements of the Research groups have had a strong focus on IPR generation -more than 50 patents- and publications -more than 45- in top Internet and Multimedia research conferences such as Sigcomm, Mobicom, NSDI, Sigmetrics, ACM Multimedia, ACM CHI, ACM Recsys, ICASSP. Our projects target both the design of complete systems and prototypes, as well as the transfer of technology to existing products.

Research and innovation creates market power by exploring and developing new markets/categories which will eventually become the next blue oceans of a company. However, converting such activities into real businesses and products that have a major impact in the real world is a non trivial task.

The Telefonica I+D Research groups have a remarkable track record in transforming mid-long term research into new market products. Recent and ongoing tech transfer efforts include designing and implementing Telefonica's video distribution network (CDN), developing the mobility algorithms that are at the core of our data analytics products, providing state-of-the-art recommendation engine for Tuenti (Spain's teenagers' leading social network) and contributing with innovative audio analysis algorithms that are part of upcoming rich communication services.

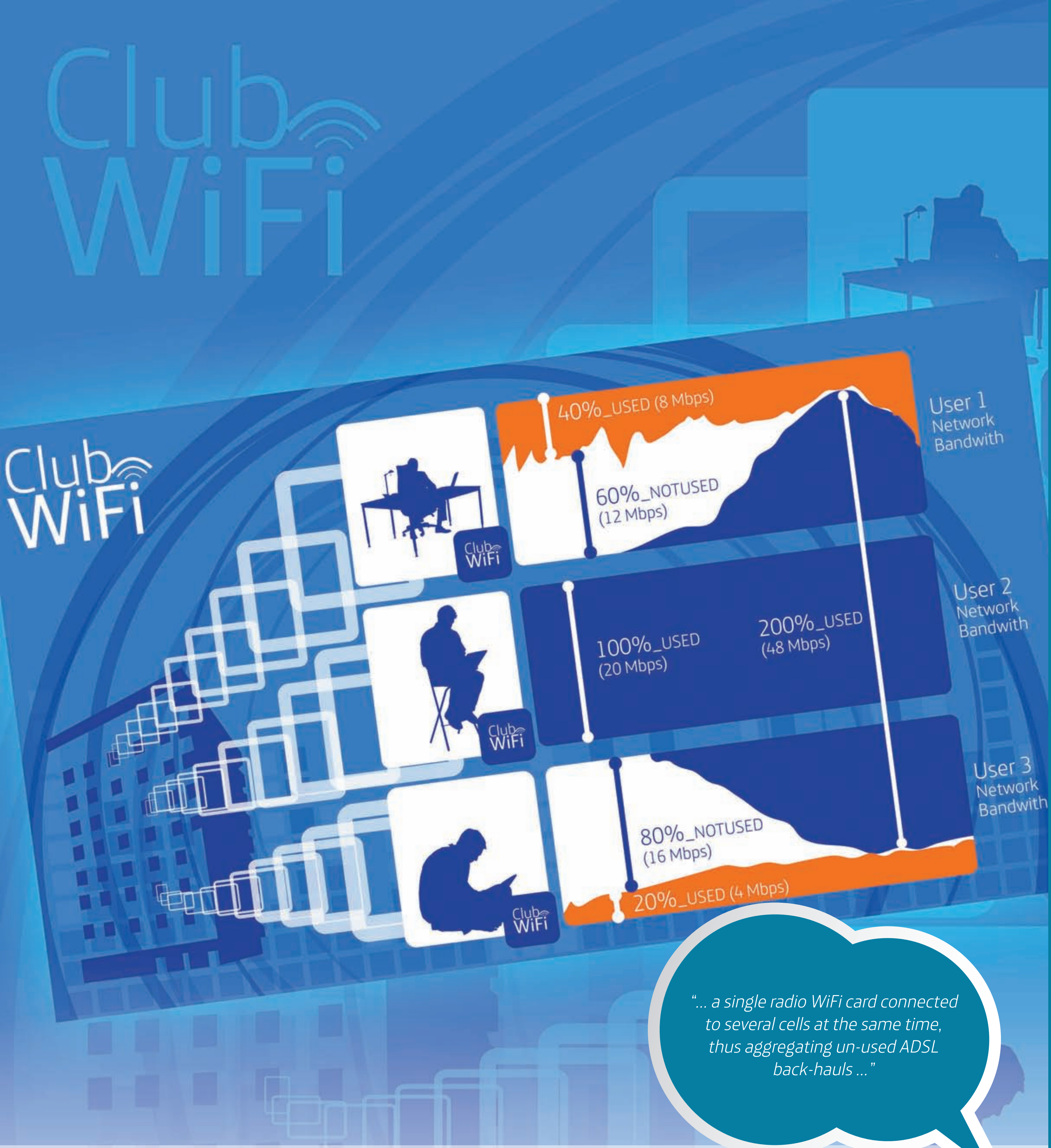
This magazine is a collection of the research projects and technologies that we have been working on over the past few years. We are confident that such innovation will enable Telefonica to maintain its leading position in the telecommunications market and be a strong player in the digital world.

Green Factsheets:
Splendorgel is an FSC Mix Certified paper, which ensures that all pulp is derived from well-managed forests and controlled sources. It is elemental chlorine free and is manufactured by an ISO 14001 certified mill.

FIXED AND WIRELESS NETWORKING

The telecommunications market has seen a dramatic change in the past 10 years. Mobile telephony has taken off and reports expect the number of mobile devices to exceed the number of personal computers in the year 2013. The volumes of content exchanged over the Internet increases by 92% every year, and is expected to stay at such levels for the next 5 years [Cisco]. More surprisingly, given the innovation on the handset front, that allows users to access Internet services on their mobile phones, a network operator reports 8000% increase in the amount of data they carry over their cellular network in just 4 years [AT&T]! To keep up with such a tremendous pace, network providers need to continuously innovate.

In the research teams at Telefonica R&D we are looking at ways to ensure the user experience while increasing network performance and minimizing operational cost.



> *"... a single radio WiFi card connected to several cells at the same time, thus aggregating un-used ADSL back-hauls ..."*

ClubWiFi: Increasing wireless bandwidth and coverage at home

In typical urban environments, residential users usually see many 802.11 gateways in range with high quality (see figure), and these gateways are commonly connected to broadband links such as ADSL and cable. At home, the users are not limited by their wireless speed (which typically ranges from 54Mbps up to 450 Mbps), but by their ADSL and cable speeds (in the range of 3 – 12 Mbps). This is particularly true in the case of uplink speeds, which rarely surpass 1 Mbps in Europe. Moreover, it is well known that these broadband links are underutilized, so they have plenty of bandwidth to spare. In this scenario, if we could simultaneously connect to all the neighboring WiFi gateways and use the *spare bandwidth* of our neighbors, we could dramatically improve our downlink and uplink speeds without impacting the quality of service of the line owners.

But aggregating the backhauls is not enough: since there could be a potential high number of users sharing their broadband links, there must be a practical way of imposing some fairness into the system. Without fairness, the perceived value of the system is severely reduced, eliminating the incentives of users to participate, and of providers to deploy it. This effectively renders the scheme infeasible.

Motivated by this problem, we developed **ClubWiFi**, a single-radio station that performs multi-gateway backhaul aggregation in a fair and distributed way, without requiring any change in the network. ClubWiFi uses a single radio WiFi card, hence working in most of the existing hardware. With ClubWiFi, the users always have their own bandwidth at home guaranteed, ensuring that the performance is always equal or better than not using ClubWiFi. Moreover, the system provides per-client fairness when several users try to share spare bandwidth of their neighbors. In this way, the users have always an incentive to participate.

We thoroughly evaluated the performance of ClubWiFi through controlled experimental tests and validated it in a deployment spanning several floors of a multistory building. We show that it achieves high aggregate throughput over the connecting gateways, and seamlessly transmits TCP traffic under realistic scenarios. In fact, Telefonica is currently piloting the ClubWiFi technology in a pre-commercial trial in several countries in Latin America.

Researchers: Alberto Lopez Toledo, Domenico Giustiniano (currently at Disney Research), (PhD student) Eduard Goma, (external), Julian David Morillo Pozo (UPC), Ian Dangerfield (NUIM), George Athanasiou (University of Thessaly)

Papers and patents:

- Giustiniano D, Goma E, Lopez Toledo A, Dangerfield I, Morillo J, Rodriguez P. Fair WLAN backhaul aggregation. In: Mobile Computing and Networking, MOBICOM '10. Chicago, IL, USA: ACM New York, NY, USA; 2010.

- Giustiniano D, Goma E, Pozo JM, Lopez Toledo A, Rodriguez P. ClubADSL: When your neighbors are your friends. In: IEEE Symposium on Computers and Communications, ISCC 2009. Sousse: IEEE; 2009.
- Giustiniano D, Goma E, Lopez Toledo A, Rodriguez P. WiSwitcher: an efficient client for managing multiple APs. In: Proceedings of the 2nd ACM SIGCOMM workshop on Programmable routers for extensible services of tomorrow, PRESTO '09. Barcelona, Spain: ACM New York, NY, USA; 2009.

Patents:

- Provisional patent: “Method for Fair Aggregation of Wireless LAN Backhauls “

"... Large amount of energy is wasted due to lax energy consumption profiles of IT devices which tend to consume close to maximum power independently of their actual workload ..."

Energy Efficient Networks

In recent years the high energy consumption of IT devices (computers, servers, networks) and supporting infrastructure (cooling systems, UPS's, etc.) has raised substantial concern for both environmental and cost control reasons. A large amount of the consumed energy goes to waste due to lack of *Energy Proportionality* in the energy consumption profile of IT devices which tend to consume close to maximum power independently of their actual workload. Making devices energy proportional is becoming a priority for component and system manufactures but this long term objective is expected to take several years until fully realized. An alternative solution that can be applied relatively fast is to permit groups of devices (server farms, network segments, etc) to behave collectively as an energy proportional ensemble, despite being made of energy un-proportional devices. The key to achieving this objective is putting some of the devices to sleep or lower power modes when the aggregate workload subsides, thus permitting the group to handle the offered load with fewer devices kept online. In Telefonica R&D we are investigating the potential of this idea in the context of two important applications/technologies: highly distributed Nano Datacenters, and broadband access networks.

Nano datacenters (NaDa) is the next step in data hosting in the content distribution paradigm. NaDa uses smart home devices that the ISP can control like smart gateways, set-top-boxes, etc to construct a managed Peer-to-Peer (P2P) network that can perform content distribution more economically than traditional centralized datacenters. Among others NaDa provides energy savings by reusing already powered on devices, avoids cooling costs, and limits the number of network hops between servers and clients.

Access networks include modems, home gateways, and DSL Access Multiplexers (DSLAMs), and are responsible for 70-80% of total network-based energy consumption.

For example, in Europe alone, it has been estimated that broadband equipment will be consuming approximately 50 TWh annually by the year 2015. This compares to the 61TWh in the US annually in data center consumption.

In Telefonica Research, we propose energy saving opportunities in broadband access networks, both on the customer side and on the ISP side. On the **user side**, the combination of continuous light traffic and lack of alternative paths condemns gateways to being powered most of the time despite having Sleep-on-Idle (SoI) capabilities. To address this, we introduce Broadband Hitch-Hiking (BHH), that takes advantage of the overlap of wireless networks to aggregate user traffic in as few gateways as possible. In current urban settings BH2 can power off 65-90% of gateways. Powering off gateways permits the remaining ones to synchronize at higher speeds due to reduced interference from having fewer active lines. Our tests reveal speedup up to 25%.

On the **ISP side**, in the distribution frame each ADSL is terminated into one of the modems belonging to a DSLAM line card. The inflexibility of current deployments makes impossible to group active lines into a subset of cards letting the remaining ones sleep. We propose introducing simple inexpensive switches at the distribution frame to enable a flexible and energy proportional management of the ISP equipment.

Overall, our results show an 80% energy savings margin in access networks. The combination of BHH and switching gets close to this margin, saving 66% on average.

Researchers: Alberto Lopez Toledo, Nikos Laoutaris, Pablo Rodriguez, Pablo Yague (interns) Eduard Goma, (external) Marco Canini (EPFL), Dejan Kostic (EPFL), Rade Stanojevic (IMDEA)

Papers:

- Goma E, Canini M, Lopez Toledo A, Laoutaris N, Kostic D, Stanojevic R, Rodriguez P, Yague P. Insomnia in the Access or How to Curb Access Network Related Energy Consumption. In: Proceedings of ACM SIGCOMM '11. Toronto, Canada: ACM New York, NY, USA; 2011
- V. Valancius, N. Laoutaris, L. Massoulie, C. Diot, P. Rodriguez, "Greening the Internet with Nano Data Centers," in Proc. of ACM CoNEXT'09.
  This paper has been slashdoted http://tech.slashdot.org/story/08/07/16/1515211/p2p-set-top-boxes-to-revolutionize-internet

Patents:

- Provisional patent: "Method for reducing energy consumption in broadband access networks (Broadband Hitch Hiking) "

“... one decade ago, researchers realized that data communication could be much more efficient if instead of passing the data unaltered from one end to the other of the communication links, sets of the data (in the form of packets) were scrambled together ...”

Network Coding

Today, the data that travels through the communication networks is like cars that travel through a highway: they enter on one end and come out of the other end (hopefully) unaltered.

However, one decade ago, researchers realized that data communication could be much more efficient if instead of passing the data unaltered from one end to the other of the communication links, sets of the data (in the form of packets) were scrambled together in one end and unscrambled at the other end. Moreover, they counter-intuitively realized that the best way of scrambling these packets were by doing it in a completely random fashion. They named this way of routing data **network coding**. That result soon drew the attention of the research community and several other works followed, opening exciting directions for the use of network coding to provide significant benefits at various stages of the network design.

In the research teams at Telefonica R&D, we have been exploring the benefits of network coding at virtually all levels of the communication stack: applications (video streaming), transport (TCP), and more low level aspects such as routing (mesh networks) or even physical layer communications (wireless security).

In particular, we set out to answer the following practical questions:

- How can we increase throughput and reliability of TCP communications in low quality wireless settings?
- Having knowledge about the network topology (e.g., in the case of multiple clients willing to consume multicast video in a controlled network), how do we to assign the available resources to achieve the optimal rate to every user?
- How can we efficiently serve a common video for users with different connection qualities and access rights in a wireless scenario?
- How can we provide incentives in peer-to-peer networks where participating users have heterogeneous requirements, while ensuring quality of service?

Motivated by these observations, we developed **CoMP**, a network coding multipath forwarding scheme that improves reliability and performance of TCP sessions in wireless mesh networks. We also analyzed the case of **degraded multicasting** that is, the case where different users require different subsets of the source content, and implemented a system architecture for network coding-based multi-resolution video streaming. We also developed a peer-to-peer video streaming protocol that provides incentives for live streaming scenarios to heterogeneous users. To this end, we designed an efficient streaming system for live video over peer-to-peer networks. Such a system accommodates large populations of heterogeneous users, behave robustly irrespective of user dynamics and ensures prescribed levels of quality of experience.

This is only a subset of the research performed by Telefonica I+D in the field of network coding. For more detailed information the reader is referred to the publication list below.

Researchers: Alberto Lopez Toledo, (interns) Steluta Gheorghiu (UPC), Fabio Soldo (UC Irvine), (external) Luisa Lima (Univ. Porto), Muriel Medard (MIT), Athina Markopolou (UC Irvine), Christina Fragouli (EPFL), Shrin Saeedi (EPFL), Xiaodong Wang (Columbia University).

Papers:

- Steluta Gheorghiu, Luisa Lima, Alberto Lopez, Joao Barros. "A Layered Network Coding Solution for Incentives in Peer-to-Peer Live Streaming", in Proc. of the IEEE International Symposium on Network Coding (NetCod 2011), Beijing, China, June 2011.
- Steluta Gheorghiu, Alberto Lopez Toledo, Pablo Rodriguez. "A Network Coding Scheme for Seamless Interaction with TCP", in Proc. of the IEEE International Symposium on Network Coding (NetCod 2011), Beijing, China, June 2011.
- Steluta Gheorghiu, Shirin Saeedi, Christina Fragouli, Alberto Lopez Toledo. "Degraded Multicasting with Network Coding over the Combination Network", in Proc. of the IEEE International Symposium on Network Coding (NetCod 2011), Beijing, China, June 2011.
- Luisa Lima, Steluta Gheorghiu, Joao Barros, Muriel Medard, Alberto Lopez Toledo. "Secure Network Coding for Multi-Resolution Wireless Video Streaming". IEEE Journal on Selected Areas in Communications, Vol. 28, Issue 3, 2010.
- Alberto Lopez Toledo and Xiaodong Wang. "Efficient multipath in wireless networks using network coding over braided meshes", International Journal of Sensor Networks 2010 - Vol. 7, No.3 pp. 176 - 188.
- Fabio Soldo, Athina Markopoulou, Alberto Lopez Toledo. "A Simple Optimization Model for Wireless Opportunistic Routing with Intra-session Network Coding". In the Proceedings of the 2010 IEEE International Symposium on Network Coding (Netcod 2010). Toronto, Canada, June 9 - 11, 2010.
- Steluta Gheorghiu, Luisa Lima, Alberto Lopez Toledo, Joao Barros, Muriel Medard "On the Performance of Network Coding in Multi-Resolution Wireless Video Streaming". In the Proceedings of the 2010 IEEE International Symposium on Network Coding (Netcod 2010). Toronto, Canada, June 9 - 11, 2010.
- Steluta Gheorghiu, Alberto Lopez Toledo, Pablo Rodriguez. "Multipath TCP with Network Coding for Wireless Mesh Networks". In the Proceedings of the IEEE International Conference on Communications (ICC 2010) - Wireless and Mobile Networking Symposium. Cape Town, South Africa. Mav 23 - 27, 2010.
- Lima, L.; Barros, J.; Medard, M.; Lopez Toledo, A., "Towards secure multiresolution network coding," IEEE Information Theory Workshop on Networking and Information Theory, 2009. ITW 2009. , vol., no., pp.125-129, 12-10 June 2009.

Patents:

- Provisional patent: “Secure Network Coding for Multiresolution Wireless video Streaming”

ONLINE SOCIAL NETWORK

Online social networks (OSNs) have seen a rapid rise the last few years and have fundamentally changed the way online content is created and consumed, and how people communicate online. Indeed online social networks like Facebook and Twitter boast hundreds of millions of users and have become indispensable. OSNs are a major source of user generated content, and are responsible for the majority of traffic being exchanged over the Internet and recently over mobile. Hence, it is imperative for a Telco to understand these systems, and the new challenges they pose. At Telefonica I+D we are looking at i) resulting system challenges, ii) the characterization and modeling of their workload, and iii) associated security implications, with the aim to use such knowledge in the better design and planning of the underlying network that makes such a service exist.

> "... twitter grew over 1300% in a single month in 2009! The issue is that by design those networks cannot be easily partitioned into smaller groups ..."

SPAR: Scaling Online Social Networks

The sheer size of today's online social networks is introducing a number of new system design challenges in scaling, management, and maintenance.

By "scaling" we mean the ability to provide a good service to millions of users, while dealing with immense popularity that could stress the system's resources. It was reported that Twitter grew over 1300% in a single month in 2009! The fundamental issue that one needs to deal with when scaling social networks is that by design those networks feature a number of interconnections that cannot be easily partitioned into smaller groups, that could for instance allow one to host different parts of the graph on different servers. The aim would be to develop a system that is able to respond to a user's question using the resources of a single machine, which in turn means that all required information to answer such a question is quickly accessible by that machine, in the best case local.

We have developed SPAR, a middleware solution that is able to transparently scale OSNs to hundreds of millions of users. SPAR stands for social network partitioning and replication. Instead of viewing the underlying interconnected data components to be a hindrance, we take advantage of the fact that such interconnections manifest in tight social communities. These tight social communities can be partitioned and by replicating nodes that lie in multiple communities, we can ensure locality of data, thus aiding in easier scaling.

We have studied SPAR using real datasets from three different OSNs - Twitter, Facebook and Orkut. SPAR provides locality semantics at the expense of moderate replication overhead.

Researchers: Georgios Siganos, Vijay Erramilli, Xiaoyuan Yang, (external) J. M. Pujol

Papers:

- Divide and Conquer: Partitioning Online Social Network, J.M. Pujol, V. Erramilli, P. Rodriguez, arXiv 0905.4918, 2009
- Scaling Online Social Networks without Pains", Pujol, J.M., Siganos, G., Erramilli, V. and Rodriguez P., Workshop on Networking Meets Databases (NetDB) in cooperation with SOSP 2009.
- The Little Engines that could: Scaling Online Social Networks, J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, P. Rodriguez, ACM Sigcomm 2010

Patents:

- Method to scaling online social networks: US Provisional patent, 2010

Press:

- Highscalability.com: http://highscalability.com/blog/2009/10/16/paper-scaling-online-social-networks-without-pains.html

"... emulate how different users post information in online social networks ..."

SONG: Social network Write Generator

Beyond the potential of Online Social Networks to enrich user's lives, they are also a tremendous source of data that can be used for social analysis, and the understanding of how people form and retain relationships. Collecting this information, however, from operational OSNs is non-trivial. Moreover, the analysis performed by analysts on one data set cannot be easily validated or continued since such data sets are very difficult to share. Having datasets to study can help fine-tuning the operations of OSNs. They can provide researchers insights on how humans communicate and interact, answering fundamental questions about the flow of information etc. And understanding such datasets can help a Telco to fine-tune their own infrastructure to handle new loads, as well as improve performance for their own customers.

We have worked on a framework, called SONG (Social Network Write Generator), that is able to generate synthetic traces of write activity on OSNs – in essence, emulate how different users post information in OSNs – like when do people update/upload data on OSNs. This can include posts/status updates, as well as uploading content. Our framework is based on a characterization study of a large Twitter data-set and the identification of the important factors that need to be accounted for. We show how one can generate traces with SONG and validate it by comparing against real data. We show that the synthetic traces are very similar to real traces, while using very few parameters.

Researchers: Vijay Erramilli, Xiaoyuan Yang

Papers:

- Explore what-if scenarios with SONG: Social Network Write Generator, Vijay Erramilli, Xiaoyuan Yang, Pablo Rodriguez, arXIv:1102.0699, under submission to journal: PhysicsA

Press:

- MIT Techreview http://www.technologyreview.com/blog/arxiv/26355/

> *"... The rise of online social media like Twitter has been accompanied by the rise of unwanted traffic on these platforms. For every success story, like their role in organizing movements in Egypt and Libya, there are examples of spam, unwanted marketing and astroturfing ..."*

Detecting Astroturfing behavior

The rise of online social media outlets like Twitter has been accompanied by the rise of unwanted traffic on these platforms. For every success story, like their role in organizing movements in Egypt and Libya, there are examples of spam, unwanted marketing and *astroturfing*, that undermine the utility of platforms like Twitter and in the worst case, can undercut real world concerns.

Astroturfing refers to the creation of an artificial "grassroots" movement to advance an agenda or promote a commercial product. By artificial, we mean an entity or collection of entities that disguise their efforts to influence public discourse by making it appear as a legitimate, independent reaction by real people.

While astroturfing has been around since the mid 80s, online astroturfing is a relatively newer phenomenon, about which little is known. In addition, astroturfing is qualitatively different from other malicious activities like spam, as the focus is more on propaganda as opposed to selling or marketing. Hence known techniques to detect spam may not work for detecting astroturfing behavior.

Our project aims to perform a thorough characterization of astroturfing profiles as gathered by the Truthy project (http://truthy.indiana.edu). We show that known methods to detect spam are marginally successful in detecting astroturfing behavior. Our study of temporal and social properties of such profiles shows that there is no single distinguishing characteristic of astroturfing profiles. Our study points towards the set of methods that can be used to uncover such profiles.

The ultimate aim of this project is to provide a service to media companies who want to receive accurate and authentic information streams and not be overwhelmed with malicious content.

Researchers: Georgios Siganos, Vijay Erramilli, (interns) D. Antoniades

Papers:

- Peeling off Astroturfing, D. Antoniades, V. Erramilli, G. Siganos, under submission

USER MODELING & MACHINE LEARNING

We live in an increasingly digitized world where our -- physical and digital -- interactions leave digital footprints. It is through the analysis of these digital footprints that we can learn and model some of the many facets that characterize users, including their tastes, personalities, social network interactions, and mobility and communication patterns. User modeling is about transforming these massive amounts of user behavioral data into meaningful customer insights for: (1) sustaining and improving the core business through strengthened business intelligence; and (2) creating new businesses, such as personalized and contextual services, or smart cities.

Below we describe the main projects in this very rich domain where machine learning and data mining techniques play a central role.

Recommender Systems

Recommender systems can be seen as a practical alternative to traditional search. They can satisfy the users' needs for relevant information without the overhead of having the users explicitly state a query. The query is therefore derived from both the user preferences and the application context. Recommender systems have proved their business value in many contexts already, ranging from e-shopping sites (e.g. Amazon) to very different settings such as television.

One of the most favored approaches to recommending is Collaborative Filtering

(CF). CF is a technique to filter or evaluate items through the opinions of other people. It makes use of peer user ratings in order to provide recommendations on the items that are unknown but may interest the target user. Collaborative Filtering-based recommender systems are typically at the core of many of today's mainstream recommendation engines (e.g. Amazon, Netflix, etc.). Despite its commercial success, it suffers from a number of limitations such as the *cold-start* problem and privacy concerns. At Telefonica Research, we are working on pushing the state-of-the-art in recommender systems by developing novel algorithms that are able to overcome some of the shortcomings of today's systems.

“... Collaborative filtering based on expert recommendations addresses many of the traditional shortcomings of filtering leveraging on massive recommendations by plane users such as noise in the user ratings, malicious attacks, the cold-start problem or privacy concerns ...”

RECOMMENDER SYSTEMS

Wisdom of the Few

In our Wisdom of the Few project, we take a different approach by using opinions or ratings form domain "experts" (i.e. individuals that we can trust to have produced thoughtful, consistent and reliable evaluations of items in a given domain). The approach, known as Expert Collaborative Filtering, is based on four steps:

1.2. Find neighbors from a reduced set of experts instead of regular users.

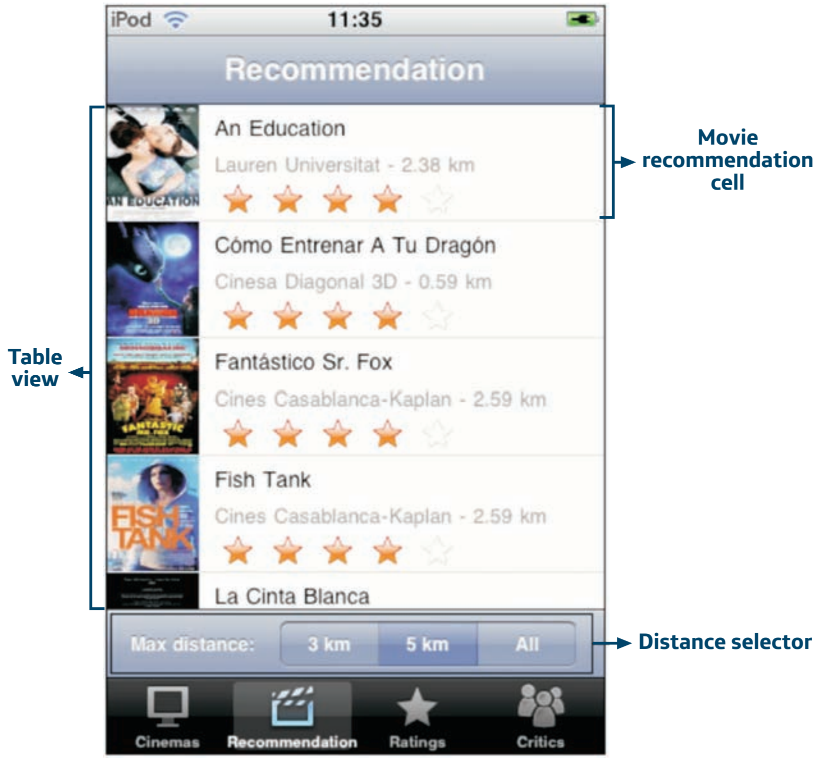
1.3. Identify domain experts with reliable ratings

1.4. For each user, compute “expert neighbors”

1.5. Compute recommendations similar to standard kNN CF

Expert CF addresses many of the traditional shortcomings of standard CF such as noise in the user ratings, malicious attacks, the cold-start problem or privacy concerns. Besides, in our user studies, we have seen that it is preferred to standard CF.

We have experimented with expert recommendations for movies using ratings from RottenTomatoes and music using ratings from Metacritic. We have a working prototype for music that uses latest web techniques such as Linked Open Data and REST services to implement a fully distributed and privacy-preserving architecture. We also have a mobile application (see Figure) based on the same approach that can recommend movies that are playing in theaters near you.



An Education
Lauren Universitat - 2.38 km

Cómo Entrenar A Tu Dragón
Cinesa Diagonal 3D - 0.59 km

Fantástico Sr. Fox
Cines Casablanca-Kaplan - 2.59 km

Fish Tank
Cines Casablanca-Kaplan - 2.59 km

La Cinta Blanca

Researchers: Xavier Amatriain and Nuria Oliver, (interns) Neal Lathia, Josep Bachs, Haewoon Kwak, Jaewook Ahn, (external) Josep Maria Pujol

Publications and Patents:

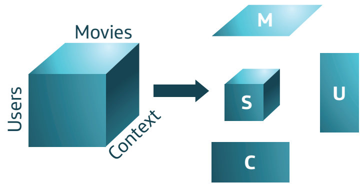
- “Geolocated Movie Recommendations based on Expert Collaborative Filtering", J. Bachs, X. Amatriain", In the 2010 ACM Recsys Conference (Demo track). Barcelona, Spain.
- "Towards Fully Distributed and Privacy-preserving Recommendations via Expert Collaborative Filtering and RESTful Linked Data", J. Ahn, X. Amatriain, In Proceedings of ACM/IEE Conference on Web Intelligenc9888e. Toronto, Canada.
- "The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web", X. Amatriain, N. Lathia, J.M. Pujol, H. Kwak, N. Oliver, in Proceedings of ACM SIGIR 09
- "Collaborative Filtering With Adaptive Information Sources", N. Lathia, X. Amatriain, J.M. Pujol, in The 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems Held in conjunction with IJCAI-09.
- Patent registered in the US. “Recommender System based on Expert Opinions”

"… understanding influence of context -location, time, social interactions, activity, weather, etc.- in determining the preference of a user in a given setting e.g. listening music, site seeing, reading books,etc. …"

RECOMMENDER SYSTEMS

Contextual Recommendations

The main aim of the Context-Aware Recommendation project is to build a compact context-aware recommender system for mobile and desktop computing devices that can integrate context data when available. Recently, the role played by context in Recommender Systems has been recognized and has contributed to increasing research efforts in the emergent area of Context-Aware



Recommender Systems (CARS).

One of the goals of this project is to increase the understanding of the influence of context (e.g. location, time, social interactions, activity, weather, etc.) in determining the preference of a user in a given setting (e.g. listening music, site seeing, reading books, etc.). Particularly we would like to answer the following questions:

- What context variables influence the preferences of users in a certain recommendation domain (e.g. movie, TV programme, music, etc)? For example does the weather influence the choice of the TV programme we view? The role of each contextual variable (e.g. time, location, activity, emotional state, social network, etc.) on users' needs is still not clearly defined.
- Which states of the context variables that are deemed to be influential determine the preference of a user? For example, are music listening patterns similar during the day and different to those during the night? Is there a particular time of the day after which we observe different listening patterns?

The main research challenge of this project is to design and develop innovative Recommender Systems algorithms and techniques that take context information into account in the recommendation process. The algorithms should:

- Offer the advantages typical of CF methods in terms of performance, computational complexity, training time, ability to handle large scale datasets and prediction time.
- Allow for the easy integration of any existing context variable and content information when available into the model.
- Handle both explicit (e.g. rating for users on items) and implicit (e.g. purchases click listening sessions that do not directly signal a strong preference for an item) taste data.

Most state-of-the-art approaches provide only limited benefits and do not adequately integrate context information into a model. Single model solutions to Context Aware Recommendation have not been proposed yet. Some CF models have been built that incorporate the temporal dimension, but they do not personalize the effects of the temporal variable but rather model global time effects.

We are currently working on single model approaches that are based on Tensor Factorization (see Figure). Tensors are multidimensional extensions of matrices. In our case, each context dimension is represented as a dimension of the tensor. A model of the interactions of the user with an item in the different context dimensions is built by factorizing the tensor.

Researchers: Alexandros Karatzoglou (Marie Curie Fellow), Xavier Amatriain and Nuria Oliver, (interns) Linas Baltrunas

Publications and Patents:

- "Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering". Karatzoglou, Alexandros, Amatriain, Xavier, Baltrunas, Linas and Oliver, Nuria. In Proceedings of fourth ACM conference on Recommender Systems, 2010
- Patent registered in the US. "Matrix Factorization based Contextual Recommendations"

> *"… "Knowing our customers" is one of Telefonica's pillars, the first global transformational program whose main objective to transform Telefonica into the best global communications company in the digital world by the end of 2012 …"*

RECOMMENDER SYSTEMS

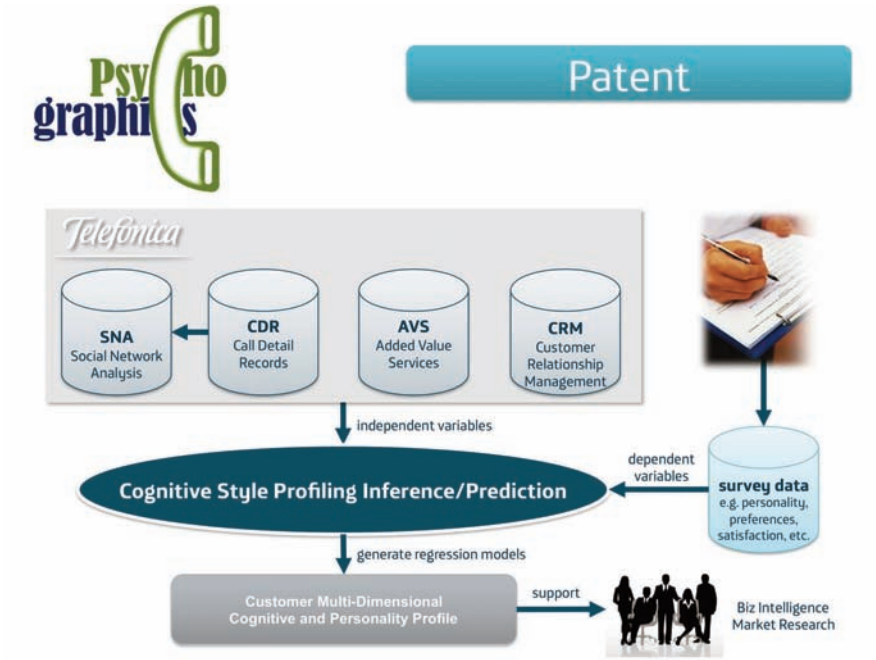
Psychographics

Knowing our customers is one of Telefonica's pillars in Bravo!, the first global transformational program whose main objective to transform Telefonica into the best global communications company in the digital world by the end of 2012. This customer knowledge opens the way to a personalized interaction with the users and the development of applications and services that satisfy existing or future customer's needs.

The main purpose of the Psychographics research project is to automatically infer the customers´ cognitive styles (cognitive and personality profiles) based on their mobile phone usage behavior as logged by the operator.

Today it is commonly accepted that our observable behavior is a consequence of internal, subjective and typically non-observable psychological features. The concept of personality has been the traditional approach to study the reasons why different people behave in different ways, even when immersed in similar or the same context. The knowledge about the customers' personality traits and psychological profile is typically obtained from market research studies, usually carried out by means of surveys. Similarly, other aspects of *cognitive styles* (e.g., customers´ recommendation roles, their technology acceptance profile, satisfaction vs. complaining profile, etc.) are also captured by these surveys. A person's cognitive profile is stable in time, and distinguishes her/him from others. The *cognitive profile* dimensions could, in principle, be linked to any kind of behavior. In practice, these cognitive dimensions allow researchers to study an individual's predisposition to have certain patterns of thought, and therefore engage in certain patterns of behavior.

This information is essential to understand behaviors that are key for a communications company, such as the ability to influence other people and the customers' behavior with respect to recommendations and complaints. Moreover, being able to construct a multi-dimensional psychological profile vector for each customer would allow to study their behavior in the social context, what is called the customer's role within her/his social network.


Psychographics
Patent
Telefónica
SNA
Social Network Analysis
CDR
Call Detail Records
AVS
Added Value Services
CRM
Customer Relationship Management
independent variables
dependent variables
Cognitive Style Profiling Inference/Prediction
survey data
e.g. personality, preferences, satisfaction, etc.
generate regression models
Customer Multi-Dimensional Cognitive and Personality Profile
support
Biz Intelligence
Market Research

The major problem in this explicit assessment approach by means of surveys is that it requires a huge amount of time and resources; it is not easily scalable; and depends on the particular scope of the study. In the Psychographics project, we propose a set of user models that predict the users' psychographic profile based on their mobile phone usage as captured by the operator, i.e. by means of Call Detail Records, variables derived from social network analysis of the call graph, information from the Customer Relationship Management, among others. Models are built from ground truth data (N=713) via statistical machine learning techniques. Our initial findings show that it is possible to predict personality profiles from mobile phone usage with up to 90% accuracy.

Researchers: Rodrigo de Oliveira, Alexandros Karatzoglou and Nuria Oliver

Publications and Patents:

- "Towards a psychographic user model from mobile phone usage", Oliveira, R., Karatzoglou, A., Concejero, P., Armenta, A. and Oliver, N. (2011) Proceedings of ACM Int. Conf. on Human Factors in Computing Systems, Work-in-progress (CHI'11), Vancouver, Canada
- Patent application registered in the US. "Customer cognitive style prediction based on mobile behavioral profiles"

"... Traditionally, urban analysis have used data obtained from surveys to characterize specific geographical areas or the behavior of groups of individuals. However, new data sources (including GPS, bluetooth, WiFi hotspots, geo-tagged resources, etc.) are becoming more relevant as traditional techniques face important limitations ..."

Smart Cities

The recent adoption of ubiquitous computing technologies by very large portions of the population has enabled the capture of large scale quantitative data about human motion. Some of the areas that directly benefit from this new source of information are urban computing and smart cities. These areas focus on improving the quality of life of an urban environment by understanding the city dynamics though the data provided by ubiquitous technologies.

Traditionally, urban analysis and the study of urban environments have used data obtained from surveys to characterize specific geographical areas or the behavior of groups of individuals. However, new data sources (including GPS, bluetooth, WiFi hotspots, geo-tagged resources, etc.) are becoming more relevant as traditional techniques face important limitations, mainly: (1) the complexity and cost of capturing survey data; (2) the lack of granularity of the data given that is typically of aggregated nature; (3) the data is static and represents a snapshot of the situation in a specific moment in time; and (4) the increasing unwillingness of individuals to provide (what they perceived to be) personal information.

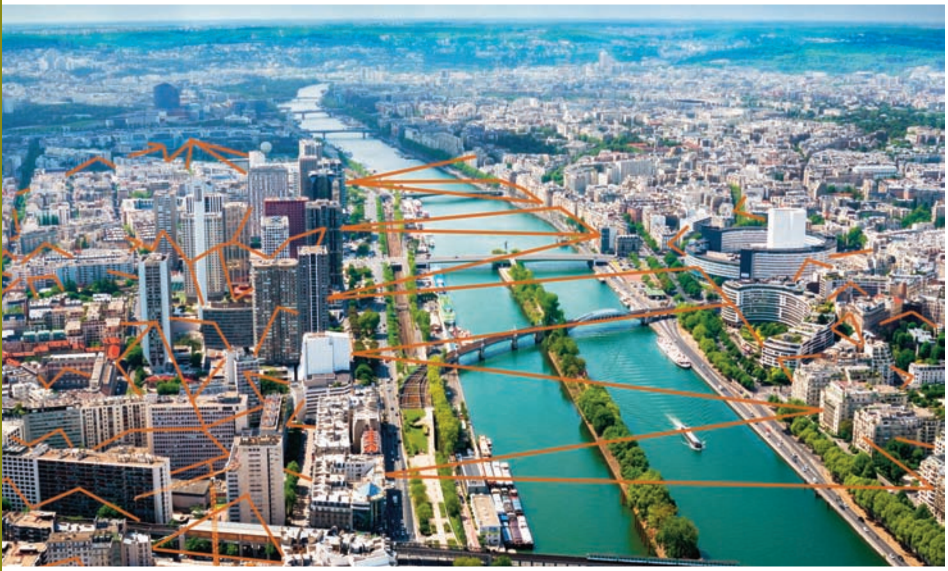
Some of the applications of smart cities and social dynamics include traffic forecasting, modeling of the spread of biological viruses, urban and transportation design and location-based services.

One of the new data sources relevant for the study of urban environments are cell phone records, as they contain a wide range of human dynamics information (ranging from mobility, to social context and social networks) that can be used to characterize individuals or geographical areas.

In this research project we use the information obtained from call detaild records to characterize and model urban landscapes in order to provide complementary approaches to traditional urban analysis techniques. The areas on which we have focussed so far include: (1) the automatic identification of dense areas; (2) the automatic segmentation of the city according to its real use; and (3) the identification of routes and mobility patterns.

The identification of areas with high density of people and/or activity is of paramount importance for e.g. urban and transport planners or emergency relief and public health officials. Urban planners can use this information to improve the public transport system by identifying dense areas that are not well covered by the current infrastructure, and determine at which specific times the service is more needed. In addition, public health officials can use the information to identify the geographical areas in which epidemics can spread faster and thus prioritize preventive and relief plans accordingly.

The spatial layout of a city has an obvious influence on the movement patterns and social behaviors found therein. Most western cities have a mixture of residential, commercial, and recreational areas connected via narrow streets, one-way avenues and a multitude of public transportation options and topographic features. Each of these areas has its own patterns of behavior which to date have only been elucidated by means of surveys and questionnaires. We have developed clustering algorithms to automatically segment the city in areas with similar behavior from the data available in anonymized and aggregated cell-phone records. Given the inherently fuzzy nature of both human behavior and urban landscapes, we propose a method to obtain robust behavioral segmentation using fuzzy clustering techniques. Using this method, only sections of the city with a given minimum similarity in their behavior will be labeled. This technique could also be applied to data obtained from other ubiquitous data sources, like geo-localized tweets, Flickr or the logs of any service that includes geo-localization.



Finally, we are working on variations of temporal association rules and Markov chains to characterize the movements in the city. This approach enables the identification of the main mobility routes and the characterization of each geographical area according to the mobility of its individuals. This knowledge can be used in a variety of domains including urban planning (proposing new public transport routes based on real movements) and efficient car-pooling.

Researchers: Enrique Frias, Vanessa Frias, Joachim Neumann, and Nuria Oliver, (interns) Victor Soto, Marcos Vieira and Jesus Virseda

Publications and Patents:

- "Robust Land Use Characterization of Urban Landscapes using Cell Phone Data", V. Soto, E. Frias-Martinez, The First Workshop on Pervasive Urban Applications in conjuntion with 9th Int. Conf. on Pervasive Computing, San Francisco, CA, 2011
- "Prediction of Socioeconomic Levels using Cell Phone Records", V. Soto, V. Frias-Martinez, J. Virseda and E. Frias-Martinez, International Conference on User Modeling, Adaptation and Personalization (UMAP), Industrial Track, Girona, Spain, 2011
- "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics", M. Vieira, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, 2nd Int. Conference on Social Computing (SocialCom2010), Minneapolis, Minnesota, USA
- "Querying Spatio-Temporal Patterns in Mobile Phone-Call Datasets", M. Vieira , E. Frias-Martinez, P. Bakalov, V. Frias-Martinez, V. Tsotras, 11th Int. Conf. On Mobile Data Management MDM2010, Kansas City, Missouri.
- Patent filed in the US. "Method for the automatic Identification of Urban Dense Ares from cell phone records"

> *"... financial crisis has revived the idea that production of goods cannot follow continuously growing trends, as suggested by Thomas R. Malthus almost two centuries ago. The shift to a service-based economy is one of the top priorities when implementing models of sustainable development. However, structural changes to the way markets are organized, in the way transportation infrastructures are used, and in the way we work and live are the hardest to achieve ..."*

SMART CITIES

Swap Relay: Leveraging Social Networks for Transportation Networks

The recent financial crisis has revived the idea that production of goods cannot follow continuously growing trends, as suggested by Thomas R. Malthus almost two centuries ago. The shift to a service-based economy is one of the top priorities when implementing models of sustainable development. However, structural changes to the way markets are organized, in the way transportation infrastructures are used, and in the way we work and live are the hardest to achieve.

In the SwapRelay research project we are designing a goods transportation network that enables its users to share items. SwapRelay leverages existing city infrastructure and people's weekly routines (obtained by means of mobile phones used as sensors) to automatically identify ways to deliver small objects to/from different parts of the city using humans as carriers.

SwapRelay is an exchange platform where people lend each other objects of small value (such as ski boots or a star-shaped screwdrivers). Objects are exchanged for free, no real money is involved. Participants earn "swap-points" that they can reuse to borrow other items. Items are insured by credit-card. Objects are transported by the owners or by carriers (owners friend or acquaintance in the social network). Carriers earn "swap points" as well. Participants in the program are not required to change their daily routines in order to deliver the items. The system is composed of a mobile component (frontend) and a backend. The frontend records the position and social interactions of the user at regular intervals and sends this information to the backend which uses this data to infer spatio-temporal mobility patterns. SwapRelay computes the optimized times at which the item can be passed onto the next person and sends reminders/alerts to the users through the mobile client.

Researchers: Mauro Cherubini and Nuria Oliver, (interns) Mengxiao Zhu, (external) Manuel Cebrian (Univ California San Diego)

Publications and Patents:

- "Exploring social networks as an infrastructure for transportation networks", M. Cherubini, M. Zhu, N. Oliver, and M. Cebrian, In Proceedings of the International School and Conference on Network Science (NetSci'10), (Boston, MA, USA), Northeastern University, May 10-14 2010.
- Patent filed in Spain. SwapRelay: Method and system to facilitate the interchange of items in a social network

Technologies for Emerging Markets

The penetration rates of cell phones in Latin America are well above the penetration rates of any other technology. This pervasiveness of mobile phones in developing economies have made them an ideal platform for providing services centered on improving local living conditions. Although these platforms have their strengths and their limitations, they provide a means to reach millions of citizens in underserved communities in both urban and rural areas.

Our research in the area of technologies for emerging markets explores two differentiated lines: (i) an Observation Approach that focuses on understanding how cell phones are currently being used in developing economies, and (ii) an Intervention Approach, with a focus on the development of new cell phone-based applications in the areas of m-learning, m-government, m-agriculture or m-health.

TECHNOLOGIES FOR EMERGING MARKETS

Observation Approach

The adoption of cell phones in emerging and developing economies has generated an unprecedented amount of data that holds information about behavioral insights, dynamics and mobility patterns in these countries. The analysis of these large datasets through machine learning techniques can help formulate usage trends characteristic in emergent economies. These analyses are of interest to both policy makers interested in the assessment of technology-based programs, and technologists and technology companies focusing on the development of personalized services for emerging economies.

Specifically, we focus on applying machine learning techniques to evaluate the impact that demographic and socio-economic indicators might have in the way people use technologies. The socioeconomic status of a population or an individual provides an understanding of its access to housing, education, health or basic services like water and electricity. In itself, it is also an indirect indicator of the purchasing power and as such a key element when personalizing the interaction with a customer, especially for marketing campaigns or offers of new products.

Traditionally, these studies have been carried out through interviews or focus groups. Our approach enhances the state-of-the-art by being able to draw conclusions automatically and from larger populations.

Researchers: Vanessa Frias, Enrique Frias and Nuria Oliver, (interns) Victor Soto

Publications and Patents:

- "Prediction of Socioeconomic Levels using Cell Phone Records", Victor Soto and Vanessa Frias-Martinez and Jesus Virseda and Enrique Frias-Martinez, In Proceedings of ACM International Conference on User Modeling, Adaptation and Personalization (UMAP), Industrial Track, Girona, Spain, 2011.
- "A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records", Vanessa Frias-Martinez, Enrique Frias-Martinez and Nuria Oliver, AAAI 2010 Spring Symposia Artificial Intelligence for Development, AI-D 2010, Stanford, USA

"... we have developed a mobile learning tool named EducaMovil that has two main components: (1) a PC tool for educational ..."

TECHNOLOGIES FOR EMERGING MARKETS

Intervention Approach

In this research line, we study the development of cellphone-based personalized applications through the use of user modeling techniques and in areas with potential social impact. This research is specially sensitive to (1) the types of platforms that subscribers have in Latin America, which are not the mobile platforms currently used in Europe (smartphones); and to (2) the personalization and adaptation of the services to each individual's interests. Additionally, this research takes input and shares techniques with the results of all the analytical work we carry out in the Observation Approach.

We are currently working in the area of mobile learning (m-learning).

While cellphones can be deployed in schools in developing countries, the greatest opportunity is to facilitate informal learning in out-of-school situations so as to complement formal schooling. In underdeveloped regions, particularly rural areas, many schools are poorly equipped or lack highly-trained teachers. In addition, school attrition can be prevalent in underdeveloped regions. In Latin America, more than 17 million children under 14 years old cannot attend school regularly because they have to work for the family in agricultural fields or households. Mobile learning thus empowers por children to balance their educational and income earning goals, by enabling them to learn anytime, anywhere, in places and times more convenient than school.

In this context, we have developed a mobile learning tool named EducaMovil that has two main components: (1) a PC tool for educational content creation and (2) a mobile game-based educational application for Java-enabled cell phones. On the PC, teachers can create the educational contents that will be shown in the mobile games. On the cell phone, the mobile game-based application consists of an open-source cell phone game where points and lives are won after correctly answering a specific quizz.

This architecture allows teachers to strictly focus on educational content creation, and uses open-source games created by game developers to provide the engaging component surrounding the educational snippets.

We are currently carrying out a longitudinal user study for a 3-month period, in collaboration with Telefonica Foundation, at a low-resource school in a peri-urban area of Lima, Peru.

Researchers: Vanessa Frias, (interns) Jesus Virseda and Andreea Molnar

Publications and Patents:

- "EducaMovil: Mobile Educational Games Made Easy", Andreea Molnar and Vanessa Frias-Martinez, In Proceedings of Ed-Media World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lisbon, Portugal, 2011

Towards Understanding Human Communication

We humans are social beings with a strong need for communication. Today, most of the communication at-a-distance is in digital form leaving in most cases a digital footprint in the form of an email sent/received, a status update/comment written on a social network, a micro-blogging post or a new entry on a phonecall database. These unprecedented amounts of data related to human communication are enabling researchers to carry out large scale studies about the nature and dynamics of communication. Our research in this area is structured around four projects:

"... Several recent studies show that human activity is strongly heterogeneous: humans act in bursts or cascades of events; most of the links are not persistent in time; and communication happens in form of group conversations. However, in most cases, the real temporal activity is aggregated over time, thus giving a static snapshot of the social interactions ..."

TOWARDS UNDERSTANDING HUMAN COMMUNICATION

Temporal Patterns and Influence in Information Spreading

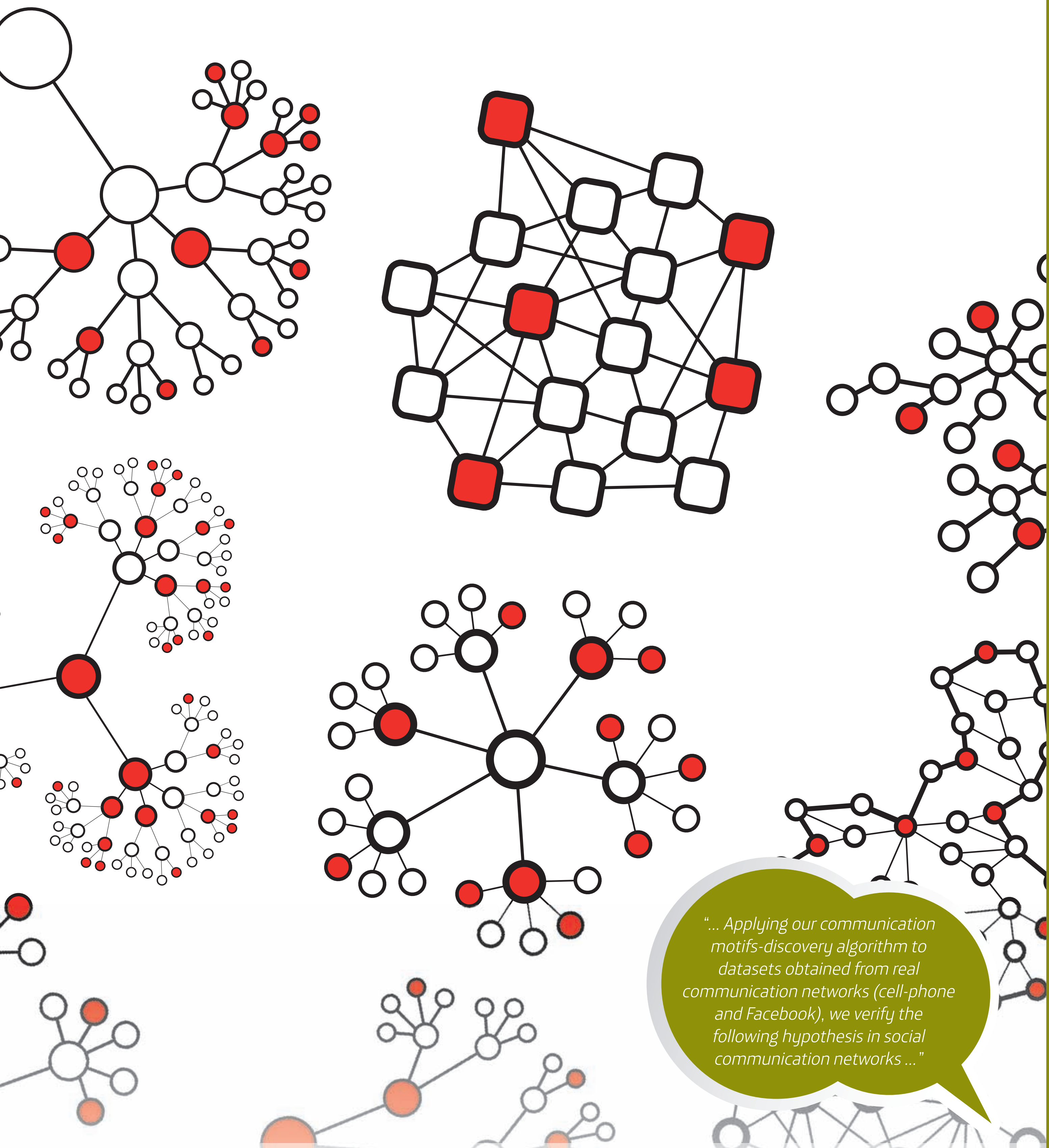
This project focuses on improving our understanding of human temporal patterns in relation with the topology of the network. Several recent studies show that human activity is strongly heterogeneous: humans act in bursts or cascades of events; most of the links are not persistent in time; and communication happens in form of group conversations. However, in most cases, the real temporal activity is aggregated over time, thus giving a static snapshot of the social interactions where links are described by static strengths that do not include information about temporal aspects of how humans interact. The results of our analyses suggest the need to consider temporal patterns of communication in the definition of a social link and in modeling of human interactions. A quantity that encompasses both the topological and the temporal patterns of human communication is the transmisibility, which represents the probability that a piece of information is transmitted from one person to another. We call this quantity the dynamic strength of a social relationship and we believe it is a key element in describing social networks and explaining information spreading phenomena. The applications of understanding and modeling dynamical social networks are multiple, including viral marketing, customer segmentation and behavioral targeting.

In collaboration with the Analytical Models innovation team at Telefonica R&D, we are analyzing topological and temporal patterns of communication to study the most influential users and characterize different strategies of viral marketing campaigns according to different products/services.

Researchers: Giovanna Miritello (PhD fellow), (analytical models innovation team) Ruben Lara, David Millan and Susana Ferreras, (external) Esteban Moro (Univ. Carlos III Madrid)

Publications and Patents:

- "The Dynamical Strength of Social Ties in Information Spreading", Miritello, G., Moro, E. and Lara, R. In Phys Review E, 83, 2011

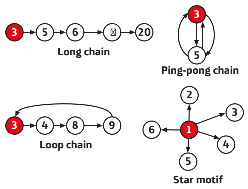


"... Applying our communication motifs-discovery algorithm to datasets obtained from real communication networks (cell-phone and Facebook), we verify the following hypothesis in social communication networks ..."

TOWARDS UNDERSTANDING HUMAN COMMUNICATION

Communication Motifs

Social networks mediate not only the relationship between entities, but also the patterns of information propagation among them and their communication behavior. In this project, we extensively study the temporal annotations (e.g., time stamps and duration) of past communications in social networks and propose two novel tools – communication motifs and maximum-flow communication motifs – to



characterize the patterns of information propagation in social networks.

Note that we are interested in the functional communication patterns of social interaction networks that not only occur frequently but also are indicative of the process of information propagation in the network. The communication motifs are therefore meant to capture recurrent patterns that appear in the collective behavior of the users in the network.

The communication motifs may reveal aspects of the principles and dynamics in the communications that take place within the network. The Figure illustrates four basic types of communication motifs with different information flow patterns: (a) *Long chain*, which is a communication graph with an extreme long list of unique participants, such as 3, 5, 6, · · · ,20 in the Figure; (b) Ping pong, which represents a pattern with very few participants that repeatedly communicate back-and-forth with each other, such as 3, 5 in the Figure; (c) Loop, where there is a loop in the communication between some of the participants – but not the ping pong-type loops, such as from 3 to 4, 8, 9, and back to 3; and (d) Star, where all the interactions start/end on a limited number of members of the graph.

In general, a communication motif involves a number of distinct participants and a combination of the basic motifs defined above. The distribution of the basic motifs in the larger motif is an indication of the topology of the social network and of how information flows within the motif and how fast it may spread.

Applying our motif-discovery algorithm to two datasets obtained from real communication networks (cell-phone and Facebook), we verify the following hypothesis in social communication networks: 1) the functional behavioral patterns of information propagation within both social networks are stable over time; 2) the patterns of information propagation in synchronous and asynchronous social networks are different and sensitive to the cost of communication; and 3) the speed and the amount of information that is propagated through a network are correlated and dependent on individual profiles.

Researchers: Nuria Oliver, (external) Qianku Zhao

Publications and Patents:

- "Communication Motifs: A tool to characterize social communications", Qiankun Zhao, Yuan Tian, Qi He, Nuria Oliver, Ruoming Jin, Wang-Chien Lee. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10, October 2010

> "... Despite the ease of communication provided by mobile phones and online social networks, people tend to devote the majority of their time to a relatively small number of ties. It is therefore resonable to imagine that individuals follow different strategies to manage the time they devote to their social relationships ..."

TOWARDS UNDERSTANDING HUMAN COMMUNICATION

Mobile Communication Strategies

Time is inelastic and people only have a limited amount of time in any given day for social interactions. Some people have many more connections than others or dedicate much more time talking on the phone/sending e-mails/using social blogs etc, than others. Moreover, people do not pay the same attention to all their relationships. Despite the ease of communication provided by mobile phones and online social networks, people tend to devote the majority of their time to a relatively small number of ties. It is therefore resonable to imagine that individuals follow different strategies to manage the time they devote to their social relationships, depending on the size of their social network and on their own rhythm and activity of communication.

Some of the questions we address in this project include an analysis and characterization of the time distribution across the social network and the identification and analysis of different communication strategies.

Researchers: Giovanna Miritello (PhD fellow) and (analytical models innovation team) Ruben Lara; (external) Esteban Moro (Univ. Carlos III Madrid), Robin Dunbar (Univ. Oxford) and Sam Roberts (Univ. Chester)

Publications and Patents:

- “Time as a limited resource: Communication strategies in mobile phone networks”, Miritello, G., Moro, E., Lara, R., Martinez, R., Belchamber, J., Roberts, S.G.B. and R.I.M. Dunbar, In Preparation

"... understanding under what conditions an edge (aka link) in a social network decay along the time ..."

TOWARDS UNDERSTANDING HUMAN COMMUNICATION

Link Decay Prediction

The project focuses on understanding under what conditions an edge in a social network at a time t is likely to decay or persist in a future time t+Dt. A social link in a communication network can be characterized by several attributes, e.g. the number/duration of phone calls, the number of common neighbors, the reciprocity, the time since the last communication, etc. Each attribute has a different predictive power that also depends on the time scale at which we observe the link. Thus, e.g., while the time since the last communication is a good predictor of the persistence/decay of the link in the following month, other variables such as the number of phone calls or the number of common neighbors might predict the persistence of the link at a larger time scale.

Researchers: Giovanna Miritello (PhD fellow) and Ruben Lara (external) Esteban Moro (Univ. Carlos III Madrid)

Publications and Patents:

- "Link Decay Problem: which features are better predictors of the persistence of a social tie?", Miritello, G., Moro, E. , Lara, R., In Preparation

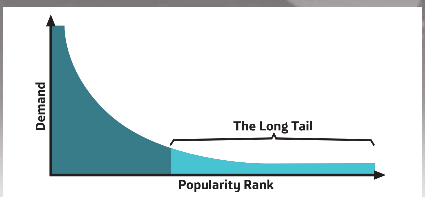
DISTRIBUTED SYSTEMS

Most applications offered in the Internet today are implemented over distributed architectures for scalability, reliability and higher performance. Redirecting clients to the closest point in the network that offers the service in question is common practice given the vastness of the digital super-highway. Service delivery is a fundamental component in the offerings of a Service Provider. At Telefonica Research we are exploring ways to make existing distributed systems' performance optimal for a worldwide infrastructure and are developing the services of the future.

Tailgate: Handling long-tail content with a little help from friends

"... content distribution technologies have witnessed many advancements over the last decade, from large Content Distribution Networks (CDNs) to P2P technologies, but most of these technologies are inadequate handling long-tail content, i.e. less popular content like the movie of your child that you recently uploaded to youtube ..."

Online content distribution technologies have witnessed many advancements over the last decade, from large Content Distribution Networks (CDNs) to P2P technologies, but most of these technologies are inadequate while handling long-tail content, i.e. less popular content like the movie of your child that you recently uploaded to youtube. CDNs find it economically infeasible to handle



such content -- the prospect of experiencing bandwidth costs and storage issues for content that may not be accessed by many people, makes handling such content impractical. P2P technologies offer a cheaper alternative but much harder to satisfy any performance guarantees.

Two recent trends have made the problem of handling such content even harder. The first is the geo-diversification of the underlying distribution architecture, that is the increasing deployment of parts of the distribution architecture at a diversity of geographic locations. This is increasingly being done to bring content closer to the end-user, decreasing latency. An additional benefit is increased reliability – safeguarding against natural disasters etc. The second is the emergence of online social networks (OSNs), where most of the content generated is long-tail. The rise of smart-phones and their increased ease of use makes generation and distribution of such content easier than ever before.

In the research grups at Telefonica R&D, we are working on a system called TailGate that exploits information from OSNs to deliver long-tail content in an efficient manner. TailGate relies on the social graph to decide where to send data, and relies on access patterns to decide when. By exploiting diurnal trends and time-zone differences, TailGate schedules content between sites such that the traffic profile is flattened -- thereby decreasing bandwidth costs that arise due to peak-based pricing schemes. TailGate also ensures content is delivered before it is likely accessed -- decreasing latency for the end-user.

Researchers: Vijay Erramilli, Nikos Laoutaris (interns) Stefano Traverso, Kevin Huguenin, Ionut Trestian

Papers:

- TailGate: Handling long tail content with a little help from friends, S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris, under submission.

Patents: US provisional to be filed

"... inter-datacenter bandwidth follows strong diurnal patterns with high peak to valley ratios that result in poor average utilization across a day ..."

Inter-Datacenter Bulk Transfers with NetStitcher

Large datacenter operators with sites at multiple locations design their inter-datacenter networks such that they can accommodate the maximum bandwidth needed during a day. At the same time, the demand for inter-datacenter bandwidth follows strong diurnal patterns with high peak to valley (i.e. high to low) ratios that result in poor average utilization across a day.

We are conducting research on how to rescue unutilized bandwidth across multiple datacenters and backbone networks. The rescued bandwidth can be used by non-real-time applications, such as backups, propagation of bulky updates, and migration of data and/or virtual machines that improve fault tolerance, end-user experience, and energy/ personnel costs, respectively. Achieving the above is non-trivial since leftover bandwidth appears at different times, for different durations, and at different places in the world.

For this purpose, we have designed, implemented, and validated NetStitcher, a system that employs a network of storage nodes to stitch together unutilized bandwidth, whenever and wherever it exists. Our system gathers information about leftover resources, uses a store-and-forward algorithm to schedule data transfers, and adapts to resource fluctuations.

We have compared *NetStitcher* with other bulk transfer mechanisms such as direct transfer, multipath forwarding, and naive store-and-forward. Our evaluation shows that NetStitcher outperforms all other mechanisms and can rescue up to five times additional datacenter bandwidth thus making it a valuable tool for datacenter providers.

We have deployed NetStitcher on Telefonica's Global Content Distribution Network to demonstrate that our solution can perform large data transfers at a much lower cost than naive end-to-end or store-and-forward schemes.

Researchers: Nikos Laoutaris, Michael Sirivianos, Xiaoyuan Yang

Papers:

- Inter-Datacenter Bulk Transfers with NetStitcher: Nikolaos Laoutaris, Michael Sirivianos, Xiaoyuan Yang and Pablo Rodriguez, ACM SIGCOMM 11
- Delay Tolerant Bulk Data Transfers on the Internet: N. Laoutaris, G. Smaragdakis, P. Rodriguez, R. Sundaram, ACM SIGMETRICS'09.
- Good Things Come to Those Who (Can) Wait or How to Handle Delay Tolerant Traffic and Make Peace on the Internet, N. Laoutaris, P. Rodriguez, ACM HotNets'08.

Patents:

- “Hermes: A Multi-hop Multi-path Store and Forward System for Bulk Transfers”, Nikolaos Laoutaris, Michael Sirivianos, Xiaoyuan Yang and Pablo Rodriguez , Telefonica Research, Spanish Patent Pending, no. P201001199
- A method for transferring tbyte sized daily tolerant bulk data using unutilized but already paid for capacity of commercial internet service providers, 0800076

> "… Amazon's Dynamo Key-Value store has emerged as a popular architecture for building Internet scale datastores. Dynamo was originally designed to fulfill the very strict set requirements of the Amazon's shopping cart …"

Key-Value Store Architectures

Amazon's Dynamo Key-Value store has emerged as a popular architecture for building Internet scale datastores. One of the key characteristics of Dynamo is the use of consistent hashing for horizontal scaling and reliability. Consistent hashing maps/partitions the keys over a single Distributed Hash Table (DHT) ring without requiring any centralized coordination. Dynamo's pioneering architecture has been used as a blueprint in many other Key Value stores that followed up, e.g. Cassandra, Voldemort, Riak to name just a few.

Dynamo was originally designed to fulfill the very strict set requirements of the Amazon's shopping cart. Current Key-Value stores, however, have relaxed and generalized the requirements in order to build data management systems that are more flexible, more generic and that can accommodate a wider range of applications and use cases. Some of the improvements proposed by Dynamo-inspired stores are the concepts of buckets as different partitioners. However, as of today, Key-Value stores still work under the constraint that one partitioner, or DHT-ring, has to apply to all data, or keys.

We propose to extend the Dynamo architecture by allowing more flexibility on how Dynamo maps/partitions keys. First, we propose an architecture where multiple concurrent DHT rings can exist, typically one per bucket. This improves the flexibility of how the datastore nodes are utilized, decouples the control plane from the data plane and allows for many domain specific optimizations. Second, we advocate for a hybrid mode where partitioners can use state to map keys to nodes. This allows for fine-grained control on how keys are mapped and allows for complex optimizations.

Researchers: Georgios Siganos, Pablo Rodriguez, (external) Demetris Antoniades, Josep M. Pujol

Papers:

- One Ring does not rule them all: Extending Dynamo's partitioning mechanism: Georgos Siganos, Demetris Antoniades, Josep M. Pujol and Pablo Rodriguez, under submission

"... a substantial amount of work has recently gone into localizing P2P traffic within an ISP in order to avoid excessive and often times unnecessary transit costs ..."

Bittorrent Locality

A substantial amount of work has recently gone into localizing P2P traffic within an ISP in order to avoid excessive and often times unnecessary transit costs. Several architectures and systems have been proposed and the initial results from specific ISPs and a few torrents have been encouraging. We are conducting large scale measurement and characterization studies in order to deepen and scale our understanding of locality and its potential. Looking at specific ISPs, we consider tens of thousands of concurrent torrents, and thus capture ISP-wide implications that cannot be appreciated by looking at only a handful of torrents. Secondly, we go beyond individual case studies and present results for the top 100 ISPs in terms of number of users represented in our dataset of up to 40K torrents involving more than 3.9M concurrent peers and more than 20M in the course of a day spread in 11K Autonomous Systems (ASes). To process these huge datasets we have developed scalable *methodologies that permit answer questions such as: "what is the minimum and the maximum transit traffic reduction across hundreds of ISPs?"*, *"what are the win-win boundaries for ISPs and their users?"*, *"what is the maximum amount of transit traffic that can be localized without requiring fine-grained control of inter-AS overlay connections?"*, "what is the impact to transit traffic from upgrades of residential broadband speeds?".

Researchers: Nikos Laoutaris, Xiaoyuan Yang, Georgos Siganos

Papers:

- Deep Diving into BitTorrent Locality, Ruben Cuevas, Nikos Laoutaris, Xiaoyuan Yang, Georgos Siganos and Pablo Rodriguez, IEEE INFOCOM 2011 mentioned in MIT technology review http://www.technologyreview.com/blog/arxiv/23924/

"... Generating realistic and updated worldwide bandwith chart infographics has been almost heroic until the moment ..."

Apollo

Apollo is a monitoring system that taps into the Bittorrent (BT) ecosystem and efficiently and discretely collects BT related performance data without requiring the cooperation of the end-users. Apollo is highly efficient and monitors hundreds of thousands of BT users on an hourly basis by utilizing a single commodity PC. Using **Apollo**, we collect a year long data capturing the download speeds of millions of BT users spread over thousands of ISPs. To demonstrate the benefits of our system we show:

(a) different traffic engineering policies, where BT users can experience more than 50% degradation of their performance across ISPs in the same country;

(b) we show that there exist significant differences in the end user perceived performance. It is not atypical to have over 100% improvement in the Download speed by switching ISPs in many countries;

(c) we show and quantify the effect of bugs and differences among BT clients. For instance, we show that there can be up to a 100% improvement between release candidates of a known BT client, and stable versions can be problematic, impacting the Download speeds for up to 20% for some ISPs;

(d) Finally, we show how Apollo's measurement have out-of-the-box applications by showing how measurements can be used to boost the download speed of BT clients.

Researchers: Georgios Siganos

Papers:

- "Comparing BitTorrent Clients in the Wild The Case of Download Speed", Iliofotou M,Siganos G., Yang X, and Rodriguez P., USENIX IPTPS in conjuction with NSDI 2010.
- "Monitoring the Bittorrent Monitors: A bird's eye view", Siganos, G., Pujol, J.M. and Rodriguez, P., Proceedings of the Passive Active Measurement conference PAM 2009.

Patents: Method of Monitoring a Bittorrent Network and Measuring Download speeds.

MULTIMEDIA ANALYSIS

Telefonica's core business is that of enabling, supporting and enhancing human communication in today's digital and multimedia-rich world. Traditional voice communications are being replaced by rich multimedia experiences that include video, music and photos; traditional home phones are being replaced by sophisticated mobile smartphones with advanced multimedia capabilities and mobile broadband access; one-to-one communication is being replaced by online social networks and micro-blogging; professional content co-exists with increasingly large amounts of user-generated content, mainly due to the pervasiveness of digital cameras and camera-phones and the popularity of online social networks. These trends are profoundly changing the nature of human communication. In this scenario, multimedia analysis tools are core technologies that are necessary to enable the communications of the future and allow the development of novel multimedia applications and services.

In the research teams at Telefonica R&D we are carrying out research on audio, image and video analysis in a variety of use cases. Below, we describe ongoing projects in the area.

"… The detection of video duplicates in a video database is one of the key technologies in multimedia management. Its main applications include storage optimization, copyright enforcement, improved web search and (d) concept tracking …"

Multimodal Video Copy Detection

The task of Content-based video copy detection (CBCD) focuses on finding video segments that are identical copies or transformations of a known video. The detection is performed using the audio and video information alone, without any embedded watermarks or having access to the original videos. Robust copy detection systems are able to find copied segments from longer videos with lengths as short as just a few seconds, which might have been affected by audio and visual transformations that altered the content to a point where it might be almost perceived as different by a user.

The detection of video duplicates in a video database is one of the key technologies in multimedia management. Its main applications include: (a) storage optimization; (b) copyright enforcement; (c) improved web search and (d) concept tracking. In storage optimization, the goal is to eliminate exact duplicate videos from a database and to replace them with links to a single copy or, alternatively, link together similar videos (called near duplicates) for fast retrieval and enhanced navigation. Copyright enforcement strives at avoiding the use and sharing of illegal copies of a copyright protected video. In the context of Web search, the goal is to increase novelty in the video search result list, by eliminating copies of the same video that may clutter the results and hinder users from finding the desired content. Finally, concept tracking in video feeds focuses on finding the relationship between segments in several video feeds, in order to understand the temporal evolution of, for example, news stories.

In the last few years several works have obtained very accurate results using only visual features. The addition of audio features brings the possibility to further improve performance and make systems more robust. Hence, effective multimodal fusion techniques become of paramount importance and are still an open research topic.

The approach that we propose differs from state-of-the-art systems in that we place equal emphasis on the detection of video copies in the audio and the video channels. Our system is able to use either modality when only one is available, and performs a robust fusion of both modalities for enhanced results. Therefore, applications that handle audio and/or video input can be built on top of the proposed framework. On the image side, we use Telefonica R&D's proprietary DART features, coupled with a scalable and fast search engine to efficiently index and retrieve large quantities of images (or hours of video).

Researchers: Xavier Anguera and Tomasz Adamek, (interns) Antonio Garzón, Daru Xu, Ehsan Younessian.

Papers and patents:

- “Multimodal Fusion for Video Copy Detection”, Xavier Anguera, Juan Manuel Barrios and Tomasz Adamek, submitted to ACM Conference on Multimedia 2011.
- “Telefonica Research at TRECVID 2010 Content-Based Copy Detection”, Ehsan Younessian, Xavier Anguera, Tomasz Adamek, Nuria Oliver and David Marimon, NIST Trecvid Workshop notebook paper
- “Telefónica Research Content-Based Copy Detection TRECVID Submission”, Xavier Anguera, Pere Obrador, Tomasz Adamek, David Marimon and Nuria Oliver, NIST Trecvid Workshop notebook paper
- “Multimodal video copy detection of social media”, Xavier Anguera, Pere Obrador and Nuria Oliver, in Proc. first SIGMM Workshop on Social Media (WSM2009) at ACM MM09
- “Audio-Based Automatic Management of Audio Commercials”, H. Duxans, D. Conejero and X. Anguera, in Proc. ICASSP 2009, Taipei, Taiwan. April 2009
- "TV advertisements detection and clustering based on acoustic information", D. Conejero and X. Anguera, in proc. International Conference on Computational Intelligence for Modelling, Control and Automation - CIMCA08, Viena, Austria, December 2008
- Patent application submitted in 2009.

> "... A challenge in multimedia content analysis is that of automatically understanding the content of audio-visual data, which would enable the development of novel and useful applications both for the end-user and the content provider ..."

Speaker-centric Audio Analysis Technologies

A challenge in multimedia content analysis is that of automatically understanding the content of audio-visual data, which would enable the development of novel and useful applications both for the end-user and the content provider. In this project we focus on the analysis of audio data, and in particular of the people whose voice is being recorded. Our aim is to create robust methods to identify *who* is speaking (referred to as speaker recognition), and *when* they are speaking (referred to as speaker diarization or tracking).

Note that current speech processing algorithms typically perform significantly better if they are queried with speech from a single speaker (speaker verification and identification) or with the different speakers already separated in order to perform speaker adaptation (speech recognition).

Direct applications for speaker recognition technology include biometrics, i.e. the verification of the person's identity by comparing his/her voice with the model from the person they claim to be; and the identification of a speaker's identity among a set of possible known speakers.
Speaker diarization algorithms focus on automatically labeling the different people that participate in a recorded conversation, indicating when each person has spoken. Such a task is usually performed without prior knowledge of the identity or the number of speakers. Current state-of-the-art systems are able to obtain good diarization error rates, but most of them are rather slow, which is an important limitation in applications that need faster than real-time processing.
Finally, speaker tracking finds when a known speaker has spoken within a long recording.

Through the collaboration of the research teams in Telefonica R&D with the University of Avignon (France) we have developed and submitted for patenting an algorithm to robustly model the speaker's unique audio characteristics and very efficiently (up to 10 times faster than state-of-the-art systems) compare them with other speaker models.

Researchers: Xavier Anguera, (interns) Esperança Movellán and (external) Jean-François Bonastre, University of Avignon, France

Papers and patents:

- "Discriminant Binary Data Representation for Speaker Recognition", Jean-François Bonastre, Xavier Anguera Miro, Pierre-Michel Bousquet, Driss Matrouf, to appear in Proc. ICASSP 2011, Prague, Check Republic.
- "Novel binary key representation for biometric speaker recognition", Xavier Anguera and Jean-François Bonastre, in Proc. Interspeech 2010, Makuhari, Japan.
- "Speaker Diarization: A Review of Recent Research", X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals., to appear in IEEE Transactions on Audio, Speech, and Language Processing (TASLP), 2011
- Patent application number FR 10/57732

"... As high-resolution digital still and video cameras become increasingly pervasive, unprecedented amounts of multimedia are being downloaded to personal hard drives and uploaded to online social networks on a daily basis. As a result, there has been an exponential increase in the overall number of photos and videos taken and shared by users ..."

Photo Storytelling and Computational Models of Media Aesthetics

Since the earliest of times, humans have been interested in recording their life experiences for future reference and for storytelling purposes. The task of recording experiences – particularly through image and video capture -- has never before been as easy as it is today. As high-resolution digital still and video cameras become increasingly pervasive, unprecedented amounts of multimedia are being downloaded to personal hard drives and uploaded to online social networks on a daily basis. As a result, there has been an exponential increase in the overall number of photos and videos taken and shared by users. This dramatic growth in the amount of digital personal media (user generated content) has led to increasingly large media libraries in local hard drives and/or online repositories, such as Flickr, Picasa Web Album or Facebook.

Unfortunately and as a consequence of this multimedia-rich world, digital information overload is becoming an increasing concern.

In Telefonica R&D's research teams, we are working to alleviate this problem. We have developed a novel multimedia storytelling system to help users both in their multimedia organization tasks and in the automatic selection of multimedia content for storytelling purposes, i.e. summarizing a collection of images or videos into a multimedia story that will be shared with other people, most likely through a social network.

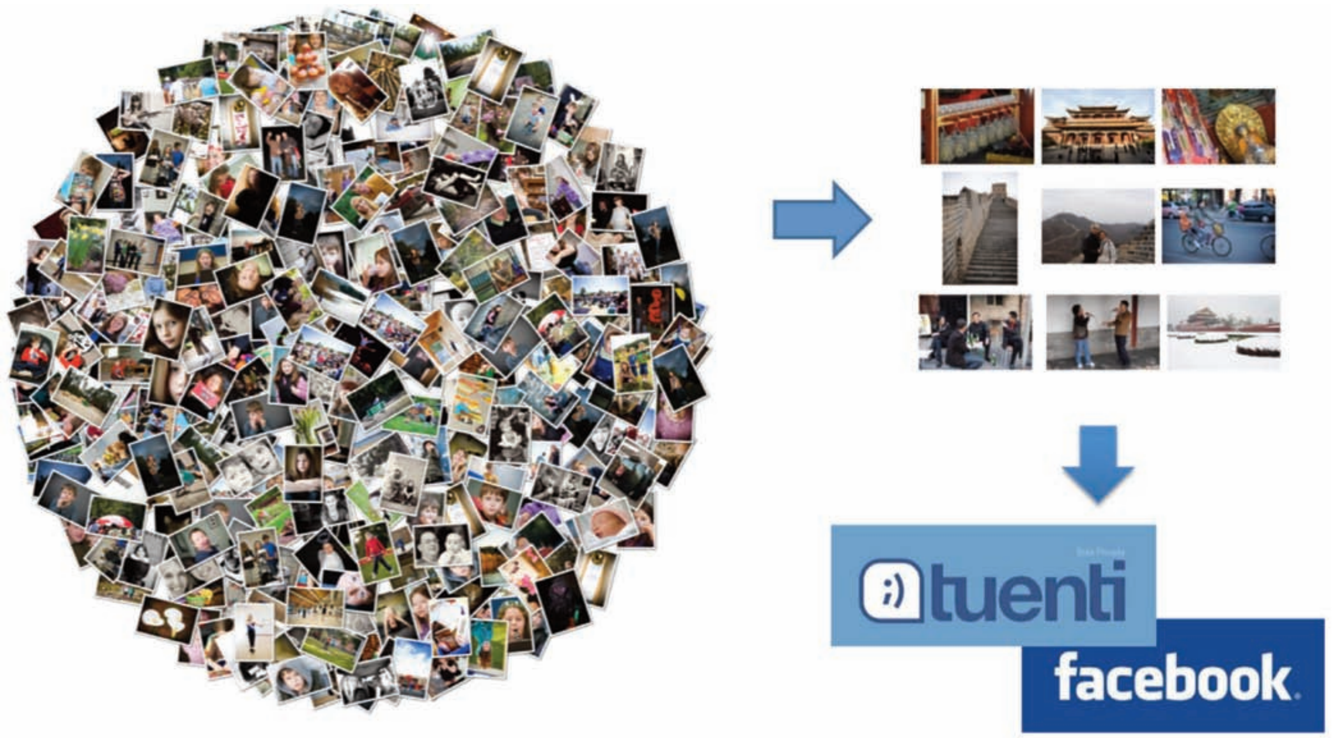
Fully automatic personal photo collection summarization for storytelling purposes is a very hard problem, since each end-user may have very different interests, tastes, photo skills, etc. In addition, meaningful and relevant photo stories require some knowledge of the social context surrounding the photos, such as who the user and the target audience are. We believe that automatic summarization algorithms should incorporate this information.

With the advent of photo and video capabilities in online social networking sites (OSN), an increasing portion of the users' social photo storytelling activities are migrating to these sites, where friends and family members update each other on their daily lives, recent events, trips or vacations. For instance, FaceBook is the largest online repository of personal photos in the world with more than 3 billion photos being uploaded monthly. Hence, there are opportunities to mine existing photo albums in OSN in order to automatically create relevant and *meaningful* photo stories for users to share online.

As opposed to some prior art in this area, neither user generated tags nor comments -- that describe the photographs, either in their local or online repositories -- are taken into account. In addition, no user interaction with the algorithms is expected. We follow an image analysis approach where both user context photos – i.e. images that are already available in the user's online social networks to which the photo stories are going to be uploaded, and the *collection* photos -- i.e., the collection of images that needs to be summarized into a story-- are analyzed using image processing algorithms. As a result, a large number of features and relevant metadata is extracted from each photo and will be used in the summarization process.

Multimedia-storytellers usually follow three steps when preparing their stories: they first choose the main *characters* in the story, next the main events to describe, and finally they select the media content (photos, videos) based on their *relevance* to the story and their *aesthetic* value. Note that one of the main contributions of this project is the design of computational models -- both regression and classification-based -- that correlate well with the human perception of the aesthetic value of images and videos.

The computational aesthetics models have been integrated into the automatic selection algorithms for multimedia storytelling, which are another important contribution of this project. We analyze the images in the collection and cluster them based on time capture, face recognition, and similarity into Characters, Acts, Scenes and Shots (following



dramaturgy and cinematography nomenclature). The photos already available in the user's social network are also analyzed in order to mirror his/her storytelling traits, mainly around the ratio of people vs non-people photos and "who" appears in those photos, helping define the story Characters.

A human-centric approach has been used in all experiments to assess the quality of both the aesthetics and storytelling algorithms, such that humans have always been the final judges of our work, either by inspecting the aesthetic quality of the media or the final story generated by our storytelling platform. In an in-depth user study, we have shown that our approach can be of help (performing as well as a human-generated summary by a professional) to users in creating a first draft of a photo album to be shared online and that the users of our system are able to generate multimedia stories with significantly less effort than if starting from scratch.

In sum, the main contributions of this project can be capitalized in two: (1) novel computational media aesthetics models for both images and videos that correlate with human perception, and (2) novel media selection algorithms that are optimized for online social network multimedia storytelling purposes.

Researchers: Pere Obrador, Rodrigo de Oliveira and Nuria Oliver, (interns) Ludwig Schmidt-Hackenberg and Anush K. Moorthy, Poonam Suryanarayan and Michele Saad.

Publications and Patents:

- "Supporting personal photo storytelling for social albums.", P. Obrador, R. de Oliveira, and N. Oliver. In Proceedings of the ACM international conference on Multimedia, ACM MM '10, pages 561--570, New York, NY, USA, 2010.
- "Audience dependent photo collection summarization.", P. Obrador, R. de Oliveira, and N. Oliver. In Proceedings of the ACM international conference on Multimedia, Grand Challenge'10, New York, NY, USA, 2010.
- "The role of image composition in image aesthetics." P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. In Proceedings of IEEE International Conference on Image Processing, pages 3185--3188. , 2010.
- "Towards computational models of the visual aesthetic appeal of consumer videos.", Moorthy, P. Obrador, and N. Oliver. In Proceedings of European Conference on Computer Vision, ECCV 2010, 6315:1--14, 2010.
- "The role of tags and image aesthetics in social image search." P. Obrador, X. Anguera, R. de Oliveira, and N. Oliver. In Proceedings of the ACM WSM '09, 1st SIGMM workshop on Social media, pages 65--72. 2009.
- Patent Application Registered in the USA. US 13098801, 2011. P. Obrador, R. de Oliveira, and N. Oliver. Automatic Storytelling for Social Albums. Telefonica.
- Patent Application Registered in Spain. P201031019, 2010. A. Moorthy, P. Obrador, and N. Oliver. Method for the classification of videos. Telefonica.
- Patent Application Registered in Spain. P201031332, 2010. P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. Image aesthetics derived from image composition low level features. Telefonica.

MIESON: Multimedia Information Extraction from Online Social Networks

The ongoing MIESON project focuses on studying methods for the analysis and exploitation of large-scale collaborative multimedia databases. This is a 3 year-long European Marie Curie IOF project. It has a strong training component, with the main objective of getting the Marie Curie fellow established as an experienced researcher in the fields of information retrieval and data analysis, with extensive knowledge and hands-on experience on the design of large-scale data-centric systems – touching on topics such as scalability, real-time data analysis, mobile computing and user experience design.

Knowledge from all those different complementary fields will be used to find sound computational solutions to current challenges in information access. A number of applications have been defined as potential targets for applying this research, including automatic content tagging or visual information aesthetics modeling, among others.

The project is divided in two different phases. The first phase, known as the outgoing phase, comprises the first 2 years of the project and will be held in the College of Information Science and Technology of the Pennsylvania State University. The focus of this phase is training on data analysis tools, data mining and machine learning techniques, high performance and cloud computing, human-computer interaction, social computing, user experience design and mobile computing. All of them are relevant to the problem of multimedia information retrieval and will enable the fellow to tackle important research problems in the area of information sciences.

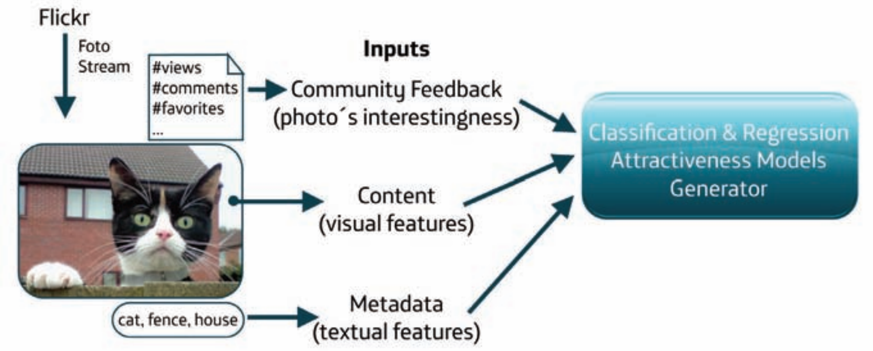
The second phase, known as the reintegration phase, will be held at the research teams in Telefonica R&D Barcelona for 1 additional year. The focus will be set on applying the results of the training and research conducted during the outgoing phase into a series of prototype systems, particularly targeting mobile platforms. It is expected that the technologies developed during the previous phase will enable the implementation of novel information access services and paradigms suitable for the specific features of mobile devices.

The multi-disciplinary and academic/industry nature of the project is important to bridge different scientific approaches, and conduct and manage transfer of knowledge and/or technology projects.

Researchers: Jose San Pedro (Marie Curie Fellow)

Publications and Patents:

- "A case for query by image and text content: searching computer help using screenshots and keywords." Tom Yeh, Brandyn White, Jose San Pedro, Boriz Katz, and Larry S. Davis. 2011. In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 775-784
- "Content Redundancy in YouTube and its Application to Video Tagging." Jose San Pedro, Stefan Siersdorfer, and Mark Sanderson. To appear in ACM Transactions on Information Systems, Volume 29, Issue 3, 2011.
- "Web-based Multimedia Information Extraction Based on Social Redundancy". Jose San Pedro, Stefan Siersdorfer, Vaiva Kalnikaite and Steve Whittaker. Book chapter to appear in "Multimedia Information Extraction" (editor Mark Maybury). 2011


Flickr
Foto Stream
#views
#comments
#favorites
...
Inputs
Community Feedback (photo´s interestingness)
Content (visual features)
Metadata (textual features)
cat, fence, house
Classification & Regression Attractiveness Models Generator

HCI AND MOBILE COMPUTING

Ten years ago, half of humanity had never made a phone call and only 20 percent had regular access to communications. Today, more than 4 billion people have a mobile phone with increasing computation and sensing capabilities. Mobile phones are the technology with widest and largest levels of penetration world-wide (and particularly in developing nations). They allow communication across remote locations and ubiquitous access to a wealth of information sources.

In the research teams at Telefonica R&D, we are working on creating differential mobile services that enhance the customers' life and leverage Telefonica's assets. We follow a user-centric approach to technological innovation with an emphasis on understanding our users, identifying existing (or future) user needs and collecting early user feedback about novel applications and services via qualitative and quantitative user studies. Some of our areas of interest include mobile information access, novel mobile context-aware applications, persuasive computing and (mobile) technologies for development – described in the User Modeling section below.

In this section, we provide an overview of recent research projects in this space.

> *"... Mobile traditionally meant on-the-move, portable, dynamic, etc. However, today an increasing number of users access the mobile Internet via their mobile phone while in non-mobile settings (i.e. at home sitting on the couch in front of the TV). This shift in the meaning of mobile is having a great effect on mobile search and mobile browsing behavior ..."*

Understanding Mobile Web usage

The term mobile is changing. Mobile traditionally meant on the move, portable, dynamic, etc. However, today an increasing number of users access the mobile Internet via their mobile phone while in non mobile settings (i.e. at home sitting on the couch in front of the TV). This shift in the meaning of mobile is having a great effect on mobile search and mobile browsing behavior. The goal of this work is to study this shift and to learn more about how it's impacting the way in which people access mobile Internet content and in particular how they use mobile search.

In order to understand more about the contexts and motivations surrounding mobile Web usage today, we carried out a 1-month user study involving 18 participants in 2011. The study employed an online diary tool where users kept track of all their mobile search and mobile Internet accesses. The online diary tool asked users what they were looking for, if mobile search was employed, what they were doing at the time of access, where they were at the time of access, etc. Following the diary study each participant engaged in a semi-structured interview were we clarified their diary entries and focused on the user intent and their frustrations while using existing mobile search and mobile Internet services. Our findings highlight significant changes in mobile Web behavior when compared to the latest previous studies and lead to a number of implications for the design of future mobile Web services, in particular related to the connection between mobile search queries and native mobile application usage.

Researchers: Karen Church (Marie Curie Fellow) and Nuria Oliver

Publications and Patents:

- "Understanding Mobile Web Use in Today's Dynamic Mobile Landscape." Church, K. and Oliver N. (2011) In Proceedings of the 13th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI'11). ACM - to appear

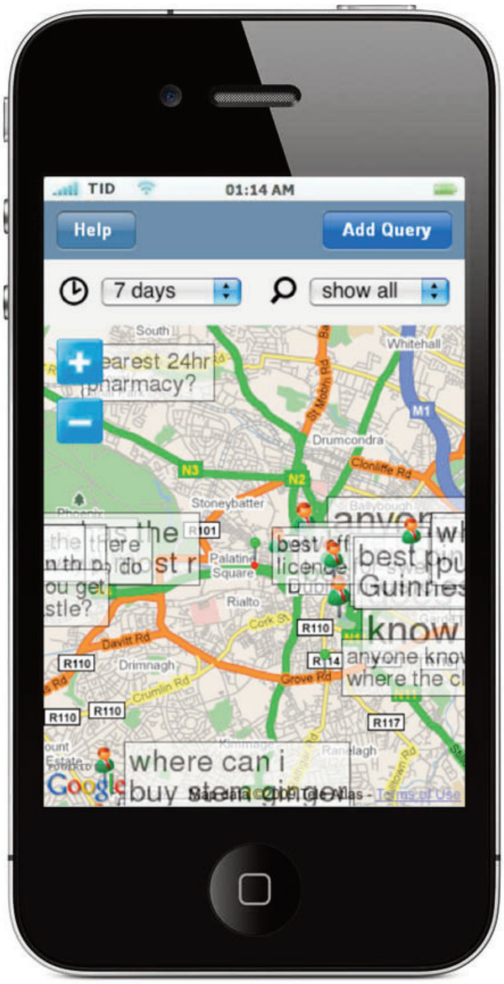
"... standard search websites have difficulties providing answers to questions such as finding information about an upcoming friend's birthday or what is the most populated club in a city after a major sporting event ..."

Social Search Browser: Exploring Social Mobile Search

The mobile Internet offers anytime, anywhere access to a wealth of information to billions of users across the globe. However, the mobile Internet represents a challenging information access platform due to the inherent limitations of mobile environments, limitations that go beyond simple screen size and network issues.

Geographical information is intimately related to mobile devices, as they are typically carried by their owners in their daily lives. Contextual information of this nature is too fined grained and too dynamic to be captured by Web technologies. Often, we seek and share local information through other communication channels and word-of-mouth appears to be the most reliable and efficient communication medium in certain information seeking tasks. For instance, standard search tools have difficulties providing answers to questions such as finding information about an upcoming friend's birthday or what is the most populated club in a city after a major sporting event. Furthermore, humans are social beings who often seek new and improved ways of sharing information with their peers. In the last few years, several research projects have attempted to exploit the social dimension of search by designing interfaces that allow users to collaboratively help each other in finding the information they need. However, these prototypes were designed with the desktop experience in mind. Conversely, prototypes that were built to improve mobile search did not exploit the social dimension of information seeking. In this regard, we are interested in understanding whether people's information needs while on-the-go could be addressed by providing a readily available connection to a user's social network. We believe that friends and family, who are trusted information sources, are likely to be able to draw on their experiences to provide interesting, valuable and relevant answers to the çuser's queries while on-the-go.

To investigate the social aspect of mobile search, we have developed SocialSearchBrowser (SSB), a proof-of-concept map-based mobile search prototype designed to enhance the search and information discovery experience of mobile users. SSB proactively displays the queries and interactions of other users in a given physical location, and taps into the social dimension to search and information access by allowing friends and other users to answer your queries while you are on-the-move. Furthermore, SSB provides novel methods for filtering the queries displayed based on the level of the friendship among users.



We carried out two live field studies of the prototype in 2009. The first study focused on an exploratory analysis of users general reactions to the prototype. The second field study focused on understanding the impact that the type of user interface has on the search and information discovery experience of mobile users. The results of these field studies have enabled us to outline a number of important implications in the design of future mobile information access applications of this nature. Below is a list of publications related to the prototype, our experiences and the field studies we carried out.

Researchers: Karen Church (Marie Curie Fellow), Joachim Neumann, Mauro Cherubini and Nuria Oliver

Publications and Patents:

- "The "Map Trap"? An evaluation of map versus text-based interfaces for location-based mobile search services" Church, K., Neumann, J., Cherubin,M. and Oliver N. (2010). In Proceedings of the ACM World Wide Web (WWW '10).
- "SocialSearchBrowser: A novel mobile search and information discovery tool." Church, K., Neumann, J., Cherubin,M., and Oliver N. (2010) In Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI '10).
- "Being Social: Research in Context-aware and Personalized Information Access @ Telefonica." Amatriain, X., Pujol, J.M. and Church, K (2010) ACM SIGIR 2010 Industry Track
- "Visual Interfaces for Improved Mobile Search." Church, K., Smyth, B., and Oliver, N. (2009) In Workshop of Visual Interfaces to the Social and Semanic Web (VISSW). Held as part of ACM IUI 2009
- "Evaluating Mobile User Experience In-The-Wild: Prototypes, Playgrounds and Contextual Experience Sampling." Church, K. and Cherubini, M. (2010) Research in the Large Workshop (held as part of UbiComp '10).

"... only 63% of patients keep complying with their medication after a year and patients take their medication only 72% of the time ..."

MoviPill

A recent review of 139 studies reporting medication adherence data showed that only 63% of patients keep complying with their medication after a year and patients take their medication only 72% of the time. Global reports presented by the World Health Organization actually reveal a more pessimistic scenario in which only 50% of the population keeps compliant to its medication prescription after one year, thus leading the US healthcare system to spend upwards of USD 290 billion in avoidable expenses every year.

In order to tackle this challenge, medical experts have tried a variety of approaches that remind patients to take their medication. Simple intervention methods such as telephonic follow-ups by the pharmacist have been shown to be effective in enhancing medication compliance and reducing the overall costs to the health provider. However, these approaches are difficult to sustain in the long-term and at a large scale.

With the pervasiveness of mobile phones and the advent of “smart packages”, automated reminding could be addressed to solve the problem of scalability. Still, previous work reports cases in which no improvement in medication compliance was observed by using reminders alone, or even where automated reminders were perceived negatively by the users.

Persuasive techniques could address this problem by shifting the focus from a human activity that we are not typically good at (i.e., remembering) to an activity that we tend to be good at and enjoy (i.e., socializing). A persuasive computing approach towards increasing levels of compliance would focus on changing the way people perceive the drug intake task. The main hypothesis would be that patients become more compliant in taking their medications when the task is not seen as an obligation, but rather as an entertaining and engaging experience.

We tested this hypothesis with **MoviPill**, a novel mobile phone-based game (see Figure) that motivates patients to be more adherent to their medication prescription by means of social competition with simple rules: more points are given to patients that take their medication closer to the prescribed time and less or negative points otherwise. The game connects all players through a social network and allows them to check how disciplined they are when compared to the other members of the social network (players).

To validate our approach, we conducted a 6-week user study with 18 elders who had to take several medications on a daily basis. Findings of this study confirmed our hypothesis. The use of MoviPill increased the elders’ compliance by 60% and the accuracy in drug intake time by: (1) 43% when applied to the entire user base and (2) 56% when the participants that were not interested in games were filtered out. This latter result highlights the importance of applying personalization in the context of persuasive computing for medication compliance.

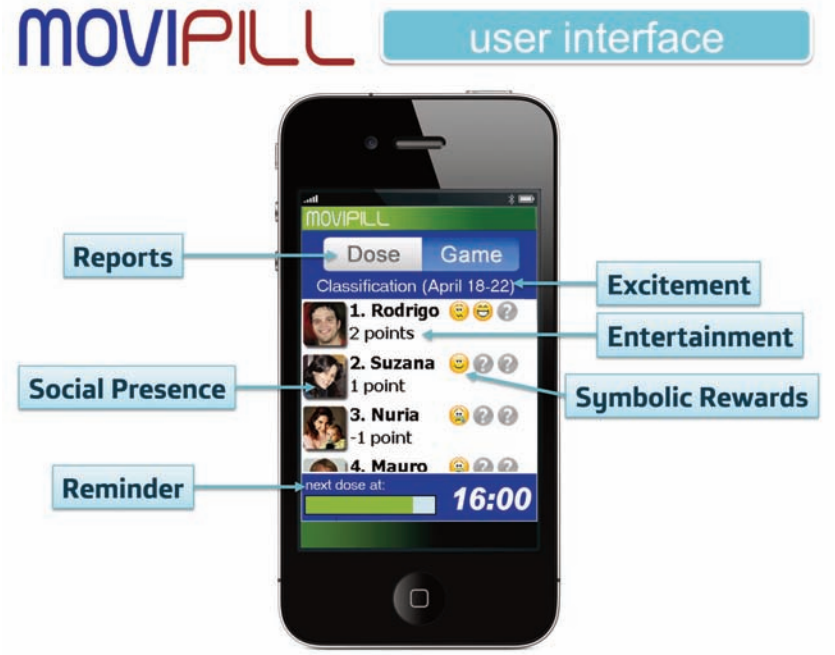
Researchers: Rodrigo de Oliveira, Mauro Cherubini and Nuria Oliver.

Publications and Patents:

- "MoviPill: Improving Medication Compliance for Elders Using a Mobile Persuasive Social Game", Oliveira, R., Cherubini, M. and Oliver, N. (2010) Proceedings of ACM Int. Conf. on Ubiquitous Computing (Ubicomp'10), Copenhaguen, Denmark, Sept 2010
- "Exploring Persuasive Techniques for Medication Compliance"
  Oliveira, R., Cherubini, M. and Oliver, N. (2010)
  WISH Workshop, Proceedings ACM Int. Conf. on Human Factors in Computing Systems (ACM CHI'10), Atlanta, US, April 2010
- Patent application registered in the US. “Improving medication compliance using persuasive computing”

Press Releases:

- Article in PSFK Future of Health (August 2nd, 2010):
  http://www.psfk.com/2010/08/future-of-health-gaming-for-health.html
- Article in La Vanguardia newspaper (July 4th , 2010):
  http://www.lavanguardia.es/premium/edicionimpresa/20100704/53957666099.html



CERMI magazine, page 28 of May/June 2011 edition.

SECURITY AND PRIVACY

Users increasingly rely on the Internet for their everyday financial transactions, purchases, communications and maintenance of their inter-personal relationships. Whenever, a resource becomes so popular and frankly indispensable for a few of us, security and privacy become important concerns. At Telefonica Research we are looking into the security and privacy implications of Online Social Networks in particular, and develop solutions that could balance the risk between feature-rich services (whose richness depends on how much the system knows about you) and the associated risk for abuse. Our work targets the management of private user information and the verification of online identities.

"... frequent and increased privacy intrusions have led to efforts to curtail the leak of personal information like cookie-blockers, privacy preserving proxie ..."

iTunes for Information

Large parts of the Internet economy are dependent on exploiting personal information of end-users. For instance, ad-revenues fuel online entities like Google, Facebook who in turn setup datacenters and purchase hardware, buy bandwidth from Telcos and invest in human capital. In return, end-users obtain services for free - - Facebook, GMail, Search etc. However there have been recent efforts to exploit more personal information, by entities like Google, Facebook and even device manufacturers like Apple, leading to increased scrutiny from privacy watchdogs as well as mainstream media.

Such frequent and increased privacy intrusions have led to efforts to curtail the leak of personal information like cookie-blockers, privacy preserving proxies etc. at the risk of hindering the business models that many of these online services operate under; for instance ads in the case of search. Hence there is a tussle between online application providers wanting to exploit more personal information and end-users who want to protect that information.

Our proposal to address this tussle is an online marketplace, where personal information can be leased to third party providers like Google etc. for adequate compensation to the end-users. By handing over control to the end-user, we increase transparency in the system, and by giving users an economic stake, we increase privacy due to awareness. The Telco acts as a middleman between users and the third party providers. We are studying this problem using a multi-disciplinary approach, using tools from economics, HCI and computer science.

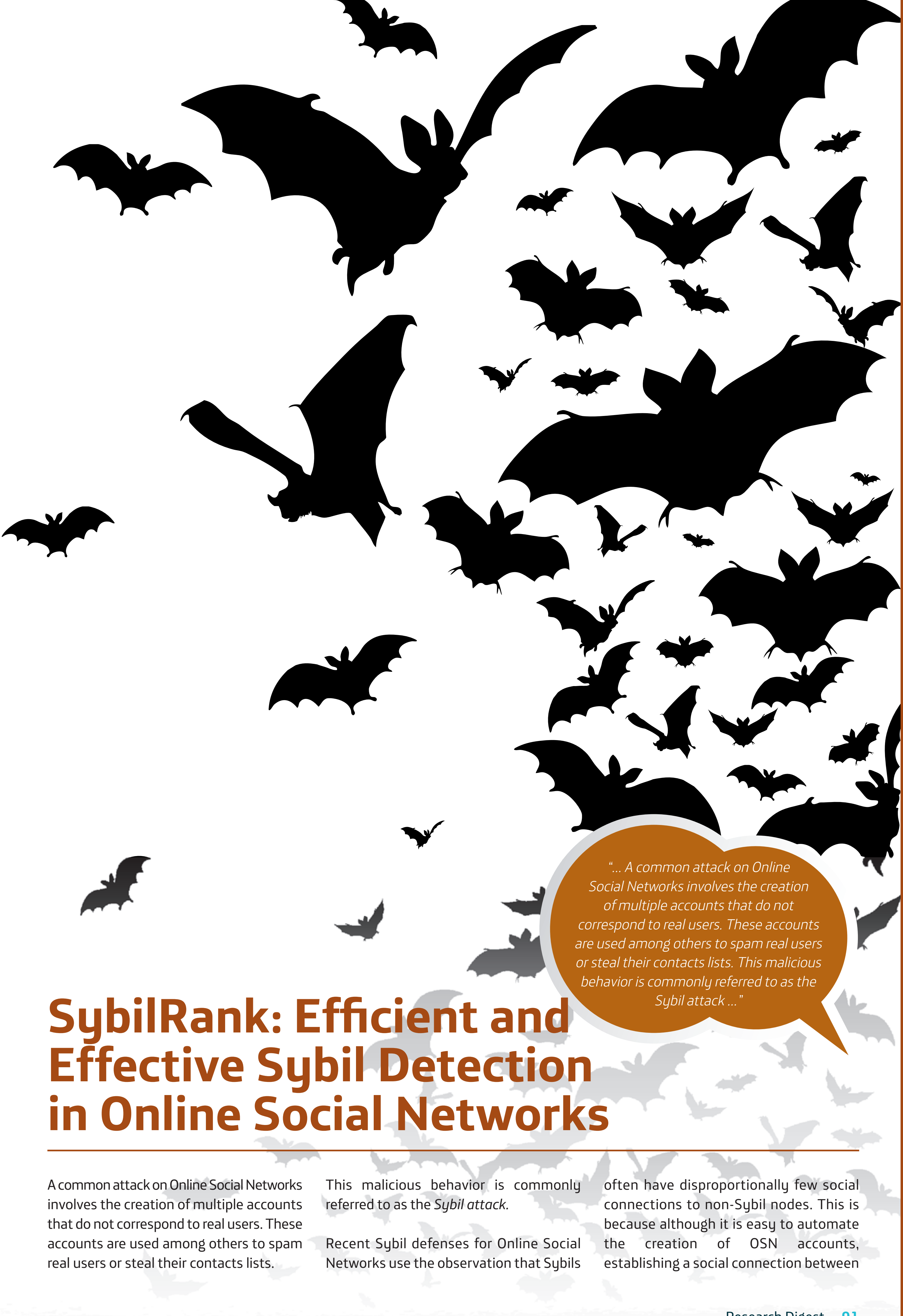
Researchers: Vijay Erramilli, Mauro Cherubini

Papers:

- Personal Information Markets: Vijay Erramilli, Pablo Rodriguez, Bala Krishnamurty, in preparation

Patents:

Transactional privacy: US provisional, to be filed.



"... A common attack on Online Social Networks involves the creation of multiple accounts that do not correspond to real users. These accounts are used among others to spam real users or steal their contacts lists. This malicious behavior is commonly referred to as the Sybil attack ..."

SybilRank: Efficient and Effective Sybil Detection in Online Social Networks

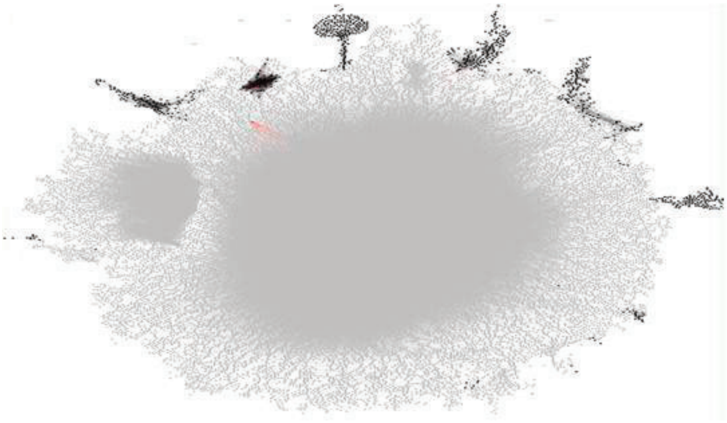
A common attack on Online Social Networks involves the creation of multiple accounts that do not correspond to real users. These accounts are used among others to spam real users or steal their contacts lists. This malicious behavior is commonly referred to as the *Sybil attack.*

Recent Sybil defenses for Online Social Networks use the observation that Sybils often have disproportionally few social connections to non-Sybil nodes. This is because although it is easy to automate the creation of OSN accounts, establishing a social connection between users implies trust that requires effort to build.

However, prior social network-based defenses are either not satisfactorily accurate or have a high computational cost.

We propose SybilRank as an effective and efficient social-network-based Sybil detection mechanism for centralized OSNs. SybilRank models efficiently computable random walks over the social graph and modifies them such that they can reliably detect Sybil (fake) accounts. Our simulation results show that SybilRank has both lower false positive and negative rates compared to state-of-the-art solutions. Furthermore, we implemented a SybilRank prototype based on Hadoop. With only 11 commodity machines on Amazon EC2, our prototype can process a graph with 160 million nodes within 33 hours.

SybilRank has been successfully deployed on Tuenti, which is the largest Online Social Network in Spain, with approximately 11 million users. Due to the diversity of reasons behind the creation of Sybils in OSNs, automated (e.g., Machine-Learning-based) approaches have thus far failed to yield high detection rates and result in numerous false positives. Consequently, Tuenti and other OSNs are currently employing a manual account verification process, driven by user reports – a process that requires a significant amount of time, and that leads to only a very small fraction of all fake accounts to be identified. We put SybilRank in the hands of Tuenti engineers and have verified that more than 90% of the what SybilRank classifies as fake accounts are indeed fake. More importantly, our tool has allowed Tuenti operations to identify 20 times as many fake accounts, as opposed to the manual process they have been following thus far.



This figure depicts a 20498 user community in the Tuenti social graph. The black dots indicate 793 nodes that SybilRank has identified as fake. The gray edges correspond to social connections between real users.

Researchers: Qiang Cao (Duke University), Michael Sirivianos, Xiaowei Yang (Duke University), Tiago Pregueiro (Tuenti)

Papers: Aiding the Detection of Fake Accounts in Large Scale Social Online Services (in submission)

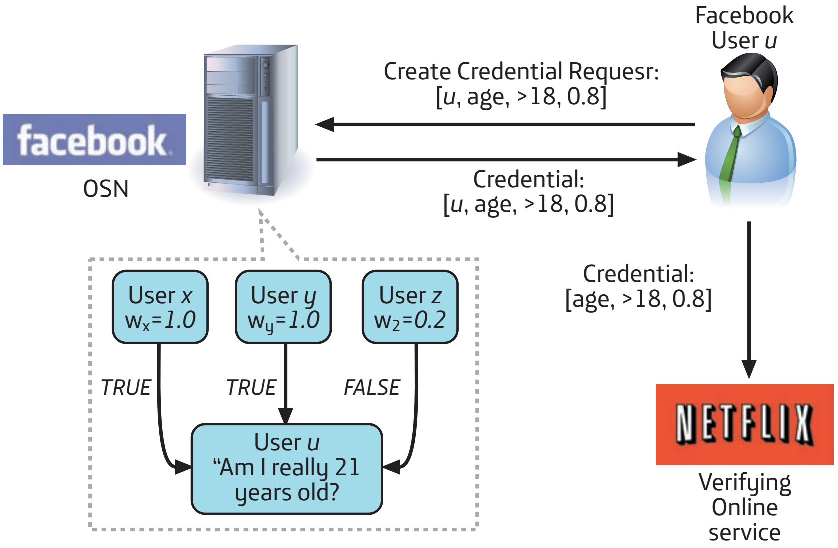
"... Assessment of the veracity of assertions that online users make about their identity attributes, such as age or profession ..."

Crowdsourcing identity credential verification

Anonymity is one of the main virtues of the Internet, as it protects privacy and enables users to express opinions more freely. However, anonymity hinders the assessment of the veracity of assertions that online users make about their identity attributes, such as age or profession. For example, a web user may easily claim that he is over 18 to gain access to inappropriate content, or an Amazon online store user may claim that he is a seasoned professional so that others place more weight on his review of a product.

We have built FaceTrust, a system that uses online social networks to provide lightweight identity credentials while preserving a user's anonymity. FaceTrust employs a ``game with a purpose'' design to elicit the opinions of the friends of a user about the user's self-claimed identity attributes, and uses attack-resistant trust inference to assign veracity scores to identity attribute assertions.

FaceTrust provides credentials, which a user can use to corroborate his assertions. We evaluate our proposal using a live Facebook deployment and simulations on a crawled social graph. The results show that our veracity scores strongly correlate with the ground truth, even when a large fraction of the social network users is dishonest and employs the Sybil attack.



Researchers: Michael Sirivianos, Kuyngbaek Kim (UC Irvine), Jian Wei Gan (TellApart) and Xiaowei Yang (Duke University)

Papers:

- M. Sirivianos, K. Kim and X. Yang. "FaceTrust: Assessing the Credibility of Online Personas via Social Networks", Usenix Workshop on Hot Topics in Security (HotSec) 2009
- Assessing the Veracity of Online Identity Assertions via Social Networks: Michael Sirivianos, Kyungbaek Kim, Xiaowei Yang, in submission to CoNEXT 2011

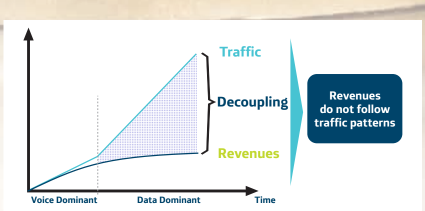
NETWORK ECONOMICS

Finally, as operators look into the future, and the offering of increasingly advanced services, a fundamental question becomes “how to price such services or even data access to ensure a viable business” and “how to share the costs with other key players of the economic ecosystem like content creators and device manufacturers”. In the research groups at Telefonica R&D we are trying to understand the different players in the Internet ecosystem and study its fundamentals.

"...Internet's basic economic rules and practices grew in an ad hoc spiral manner from the convolution of economic tensions, technology, user demand trends, policy & regulation..."

Pricing policies for end customers & ISP inter-connection economics

Like so many other aspects of the Internet, it's basic economic rules and practices grew in an ad hoc spiral manner from the convolution of economic tensions, technology, user demand trends, policy & regulation. All of the above have left imprinted marks on de facto business practices like the flat pricing of residential broadband access, peak rate pricing for network transit services (95-percentile



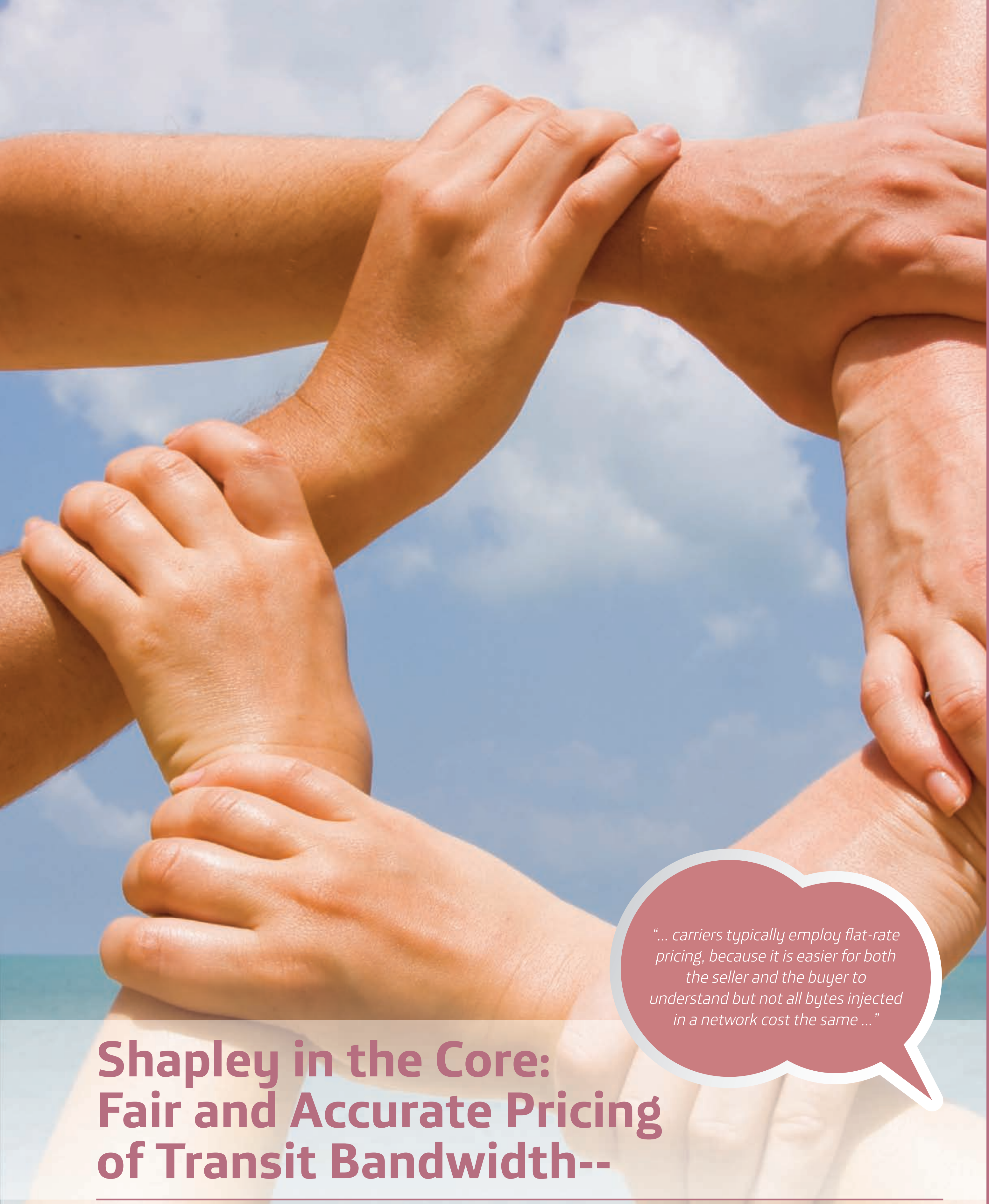
pricing rule), and the charge-free peering between equally sized networks for exchanging their direct traffic and avoiding paying their transit provider. Whereas there exists a huge body of traditional economics's literature on the above issues it is usually limited by the lack of publicly available data on the operations and costs of networks or by computational challenges due to huge volume, when such data are indeed available. In Telefonica we are taking a data/computation driven approach to the study of several research questions in the general area of network economics. Current studies include, pricing for residential broadband; inter-connection economics between access ISPs, transit ISPs, and content creators; secure multi-later payment schemes for bulk flows.

Papers:

- On Economic Heavy Hitters: Shapley value analysis of 95th-percentile pricing, R. Stanojevic, N. Laoutaris, P. Rodriguez, ACM IMC'10. J. Trade & Cap: A Customer-Managed, Market-Based System for Trading Bandwidth Allowances at a Shared Link, Londoño, A. Bestavros, N. Laoutaris, USENIX NetEcon'10.
- Home is where the (Fast) Internet is: Flat-rate Compatible Incentives for Reducing Peak Load, P. Chhabra, N. Laoutaris, P. Rodriguez, and R. Sundaram, ACM SIGCOMM Workshop on Home Networking, 2010.

Patents

P0800063 - OFFERING INCENTIVES UNDER A FLAT RATE CHARGING SCHEME



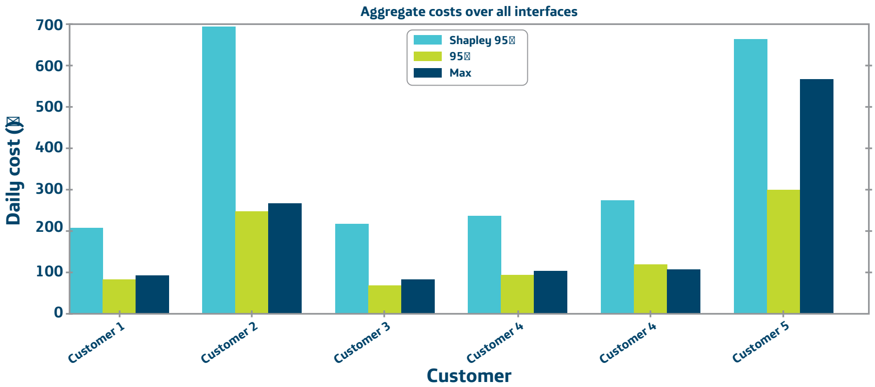
"... carriers typically employ flat-rate pricing, because it is easier for both the seller and the buyer to understand but not all bytes injected in a network cost the same ..."

Shapley in the Core: Fair and Accurate Pricing of Transit Bandwidth--

Not all bytes injected in a network cost the same. For example, volumes sent during peak demand hours impose high costs on carriers, such as Telefonica's International Wholesale Services (TIWS) . This is because they yield additional over-dimensioning investment, i.e., they force a carrier to build even higher capacity in its networks at a high cost. However, carriers typically (if not always) employ flat-rate pricing, because it is easier for both the seller and the buyer to understand.

This research aims at using i) TIWS bandwidth resource monitoring tools to obtain detailed per-customer and per-flow data, and ii) the Shapley value framework to derive fair pricings per customer and per type of flow/service. We will use this analysis to determine pricing discrepancies, and we will suggest new simple pricing models that better approximate the Shapley value in the TIWS backbone.

Researchers: Laszlo, Michael, and Niko



This figure depicts the daily costs that 6 customers incure on TIWS infrastructure. We observe substantial differences between the Shapley-based and the maximum- or 95th-percentile-based allocations of costs among customers.

What would you do if you had the **opportunity** to **shape** the **digital life** of at least **270 million people**?
Whatever it is,
you can do it
@tefdigital.
www.tid.es