Simulation Modelling Practice and Theory 93 (2019) 42–64

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

A survey on cost-effective context-aware distribution of social data streams over energy-efficient data centres

Irene Kilanioti*[,a], Alejandro Fernández-Montes[b], Damián Fernández-Cerero[b], Christos Mettouris[a], Valentina Nejkovic[c], Rabih Bashroush[d], George A. Papadopoulos[a]

[a] Department of Computer Science, University of Cyprus, Cyprus
[b] Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Spain
[c] Faculty of Electronic Engineering, University of Nis, Serbia
[d] University of East London, UK

ARTICLE INFO

Keywords:
Social Data Streams
Social Media
Social Networks
Contextaware
Content Distribution
Multimedia Content
Energy-efficient
Data Centers
5G
Infrastructure
Cost-effective

ABSTRACT

Social media have emerged in the last decade as a viable and ubiquitous means of communication. The ease of user content generation within these platforms, e.g. check-in information, multimedia data, etc., along with the proliferation of Global Positioning System (GPS)-enabled, always-connected capture devices lead to data streams of unprecedented amount and a radical change in information sharing. Social data streams raise a variety of practical challenges, including derivation of real-time meaningful insights from effectively gathered social information, as well as a paradigm shift for content distribution with the leverage of contextual data associated with user preferences, geographical characteristics and devices in general. In this article we present a comprehensive survey that outlines the state-of-the-art situation and organizes challenges concerning social media streams and the infrastructure of the data centres supporting the efficient access to data streams in terms of content distribution, data diffusion, data replication, energy efficiency and network infrastructure. We systematize the existing literature and proceed to identify and analyse the main research points and industrial efforts in the area as far as modelling, simulation and performance evaluation are concerned.

1. Introduction

1.1. Characteristics of social data streams

Social networks, media and platforms enable communication, exchange, business and knowledge acquisition and social network users connect with each other with the purpose of sharing content. Social data is the information that social media users share, e.g. check-in information, multimedia data, tags, annotations, and likes, and may include metadata such as the user's location, language spoken, biographical data and shared links, whereas 'streams' denotes that we do not refer to static datasets, but rather to dynamic information generated and transmitted over the Online Social Network (OSN).

Formally, an OSN is depicted by a directed graph $G = (V, E)$, where V is the set of the vertices of the graph representing the nodes

* Corresponding author.
E-mail addresses: koilanioti@dbs.ifi.lmu.de (I. Kilanioti), afdez@us.es (A. Fernández-Montes), damiancerero@us.es (D. Fernández-Cerero), mettour@cs.ucy.ac.cy (C. Mettouris), valentina@elfak.ni.ac.rs (V. Nejkovic), r.bashroush@qub.ac.uk (R. Bashroush), george@cs.ucy.ac.cy (G.A. Papadopoulos).

https://doi.org/10.1016/j.simpat.2018.11.004
Received 23 June 2018; Received in revised form 7 November 2018; Accepted 8 November 2018
Available online 09 November 2018
1569-190X/

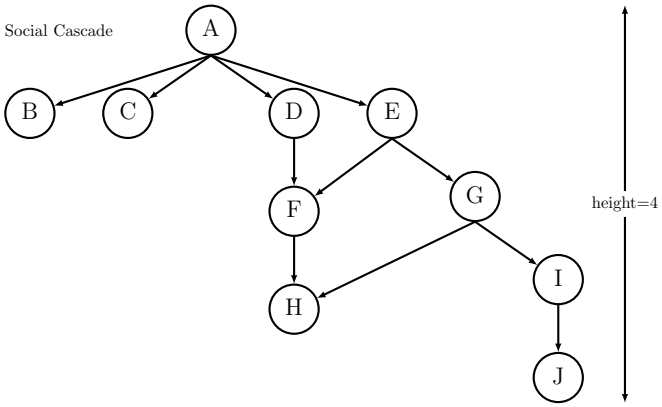

Figure 1. The evolution of a social cascade in Twitter.

of the network and *E* are the edges between them, denoting various relationships among the edges of the graph [58]. The semantics of these edges vary, and their interpretation is expanded for various OSNs from personal acquaintance, to common interests, microblogging services or business contact. As far as the directionality of the edges of the social graph is concerned, it is associated with the concept of the OSN: for Facebook, an edge denotes mutual friendship between the endpoints of a link, for Twitter, if the edge between A and B points at B, A's posts (tweets) appear in B's main Twitter page, etc. A social node centrality is indicative of the importance of a node within a social network. It is given in terms of a real-valued function on the vertices of a graph, where the values produced are expected to provide a ranking which identifies the most important nodes [29,30].

In his classic work [130] Rogers defines *information diffusion* as the process in which an innovation is communicated through certain channels over time among the members of a social system. In this context, the innovation is defined as the first spread of an information from an originator. A *social cascade* is a specific case of information diffusion and practically occurs within an OSN, when a piece of information is extensively retransmitted after its initial publication from a user. Cascades can be represented as rooted directed trees where the initiator of the cascade is the root of the tree [24] and the length of the cascade is the height of the resulting tree. Each vertex in the cascade tree can have the information of the user, and the identity of the item replicated in the cascade. Fig. 1 depicts an example of the evolution of a social cascade in a directed graph. The cascade follows the arrows' direction. For example, in Twitter, B, C, D, E are followers of A, whereas the adopters of a new information piece could be the nodes, that after having been exposed in a video link, they retransmit it, contributing remarkably to Internet traffic [8].

1.2. Challenges for distribution of social data streams

Our survey focuses on the challenge of enabling better provisioning of social media data based on the context of users accessing these resources. Distributing social data streams largely depends on the exploitation of usage patterns found in OSNs, and can be improved either through the selective prefetching of content **(cost-effectiveness)** or through the strategic placement/ selection of the employed infrastructure **(energy-efficiency)**. The cost of scaling such content might be the number of replicas needed for a specific source or it may take into account the optimal use of memory and processing time of a social-aware built system. Optimization of energy efficiency for data centres that support social data interaction and analysis includes tasks such as data growth, data centre federation and CDN-load-balancing at data centre level. In our taxonomy (Fig. 2), pillars associated with cost-effectiveness include Context-aware Computing, Content/Information Diffusion Models and Content Distribution challenges, whereas Software for Infrastructure Efficiency is associated with energy-efficiency. This taxonomy includes solutions or approaches to the 'Challenges for Distribution of Social Data Streams'. These solutions or approaches require enough effort, hence they can also be considered as a challenge for the research community.

Context-aware computing: Application of social contextual information, such as profiles, images, videos, biometrical, geolocation data and local data, in situations where conventional bandwidth-intensive content scaling is infeasible could largely facilitate the spreading of information, the identification of potential information sources, as well as a paradigm shift in the way users access and control their personal data. Application of social contextual information, such as profiles, images, videos, biometrical, geolocation data and local data, in situations where conventional bandwidth-intensive content scaling is infeasible could largely facilitate the spreading of information, the identification of potential information sources, as well as a paradigm shift in the way users access and control their personal data. User-generated multimedia content is especially difficult due to its long tail nature, with each item probably not popular enough to be replicated in a global scale, but with the long-tail altogether getting sufficient accesses [9]. Social analysis tasks interweaved with context-aware computing could pave the ground for preactive caching mechanisms in the framework of a content delivery infrastructure of streaming providers.

Software for infrastructure efficiency: The industry has made several efforts to address challenges associated with optimization of energy efficiency for data centres that support social data interaction and analysis [23,108,146] such as data growth, isolation, real-time interactions, data centre federation and CDN-load-balancing at data centre level, but usually lacks from focusing on energy

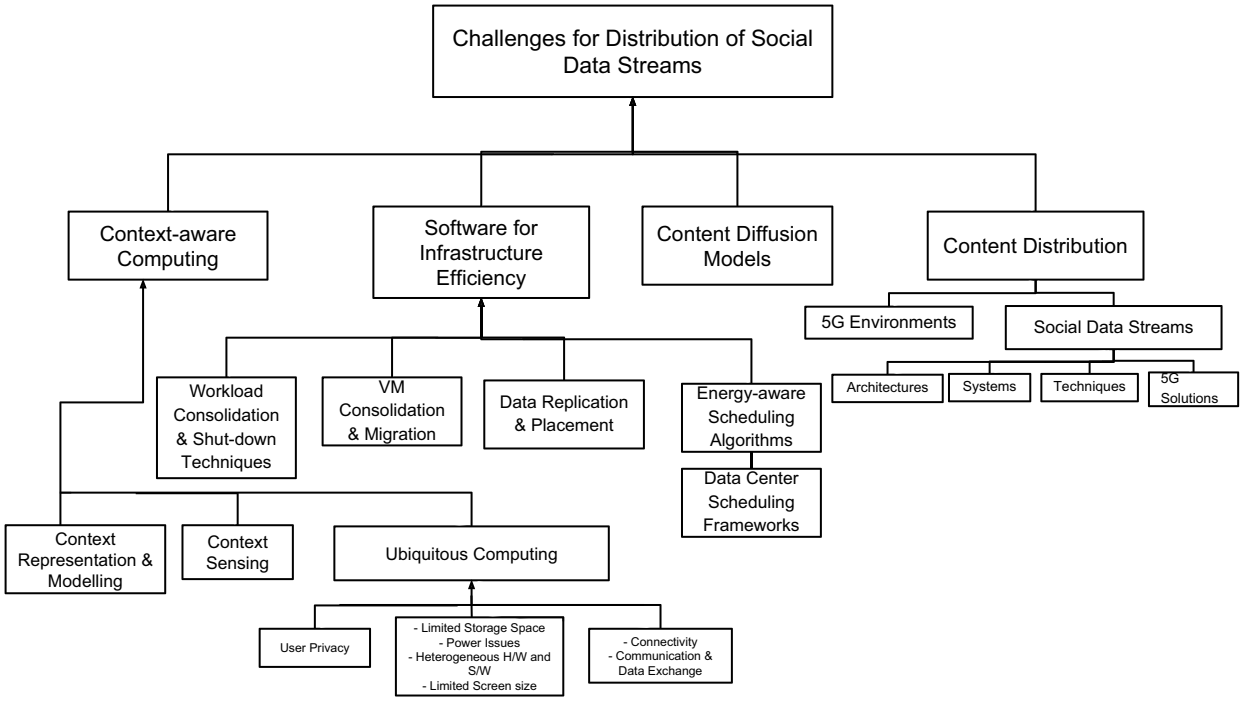

Figure 2. Taxonomy of challenges for distribution of social data streams.

consumption of the employed infrastructures. The challenges in the area of energy-efficient data-centres include workload consolidation and shut-down techniques, Virtual Machines (VMs) consolidation and migration, data replication and placement, and energy-aware scheduling algorithms.

Content/information diffusion models: Prevalence of OSNs has transformed the landscape of content exchange. Popularity of relatively data heavy multimedia user generated content (UGC) has also risen [82], resulting in data deluge across all media platforms [159], [109,121]. Measurement studies, such as [40], attribute the recent increases in HTTP traffic to the extended use of OSNs [32,36,58]. Elaborate data manipulation presupposes coping with the size of social graphs with billions of nodes and edges [158]. Facebook, for example, reported that had 1.47 billion daily active users on average for June 2018 and 2.23 billion monthly active users as of June 30, 2018 [63]. Its custom built-in data warehouse and analytics infrastructure [41] has to apply ad-hoc queries and custom MapReduce jobs [49] in a continuous basis on over half a petabyte of new data every 24 hours for the creation of meaningful aggregations and analyses. It is also acknowledged that a large proportion of bandwidth-intensive media is distributed via reposted OSN links, contributing significantly to Internet traffic [8,33]. These challenges are closely associated with the Content/Information Diffusion Models used to represent the diffusion of information over OSNs and facilitate relevant algorithmic solutions (Fig. 2).

Content distribution: The delivery infrastructure of video operators is made up of scattered geo-distributed servers, which with specific cache selection mechanisms direct users to the closest servers hosting the requested data. Transmission Control Protocol (TCP), however, is subject to delay jitter and throughput variations and clients are required to preload a playout buffer before staring the video playback [96]. Thus, the quality of experience (QoE) of media platform users is primarily determined by stalling effects on application layer. Cache server selection is also highly Internet Service Provider (ISP)-specific for the YouTube case with geographical proximity not being the primary criterion and DNS level redirections for load-balancing purposes occurring quite frequently and substantially contributing to the initial startup delay of the playback. Several network-level and client-level approaches are focused on the detection of such interruptions, that negatively affect the user experience [80]. With the growing popularity of OSNs and the increased traffic due to outspread of information via the latter, the improvement of user experience through scaling bandwidth-demanding content largely depends on the exploitation of usage patterns and geolocation data associated with OSNs. These challenges are closely associated with the Architectures, Systems and Techniques within the 5G infrastructure employed.

Below we discuss some key factors contributing to the problem of diffusion of bandwidth-intensive media content over OSNs.

1.2.1. Large-scale datasets

In order to harness the power of social networks diffusion over (Content Delivery Network) CDN infrastructure, the key areas of interest that need to be explored include the large size of the graphs, and also the fact that diffusion of links is multiplied through dissemination over sites like YouTube, and amplified by the proliferation of smartphones and cheap broadband connections. The amount of information in OSNs is an obstacle, since elaborate manipulation of the data may be needed. An open problem is the efficient handling of graphs with billions of nodes and edges.

The desired scaling property refers to the fact that the throughput of the proposed approaches should remain unchanged with the

increase in the data input size, such as the large datasets that social graphs comprise and the social cascades phenomena that amplify the situation. Cost of scaling such content can be expressed in different ways. For instance, it may be matched with the number of replicas needed for a specific source. Future experimentations may take into account the optimal use of memory and processing time of a OSN-aware built system.

Internet of Things (IoT) is a global infrastructure that interconnects things based on interoperable information and communication technologies, and through identification, data capture, processing and communication capabilities enables advanced services [85]. Things are objects of the physical world (physical things, such as devices, vehicles, buildings, living or inanimate objects augmented with sensors) or the information world (virtual things), capable of being identified and integrated into communication networks. It is estimated that the number of Internet-connected devices has surpassed the human population in 2010 and that there will be about 50 million devices by 2020 [61], thus, the still undergoing significant innovation IoT is expected to generate massive amounts of data from diverse locations, that will need to be collected, indexed, stored and analysed.

1.2.2. OSN evolution

Existent works examine valuable insights into the dynamic world by posing queries on an evolving sequence of social graphs (e.g. [128]) and time evolving graphs tend to be increasingly used as a paradigm also for the emerging area of OSNs ([62]). However, the ability to scalably process queries concerning the information diffusion remains to a great extent unstudied. With the exception of sporadic works on specialized problems, such as that of inference of dynamic networks based on information diffusion data [129], we are not aware of relative studies on the information diffusion through OSNs under the prism of graphs dynamicity.

1.2.3. 5G approaches

The demand for high-speed data applications that has risen in recent decade lead to development of Fifth Generation Wireless (5G). Development of efficient mechanisms for supporting mobile multimedia and data services is prerequisite for 5G networks. Real bottleneck of todays" mobile networks is access radio network and the backhaul. Caching in the intermediate nodes, servers, gateways, routers, and mobile users" devices can reduce doubled transmission from content providers and core mobile networks.

Known caching techniques that can be used within 5G are: content distribution network, information-centric networks, content-centric networking, http web caching, evolved packet core caching, radio access network caching, device to device caching, proactive caching, predictive caching, cooperative caching [5]. Those techniques are using different algorithms and models. Analysis presented in [5] has shown that the deployment of those caching techniques in mobile network can reduce redundant traffic in backhaul, minimize the traffic load, increase the transfer rate in mobile network and reduce the latency. Correlation of several caching methods and procedures could result in improving network performance and obtaining better results.

On the other side well known bottleneck that 5G brings is complex heterogeneity of the network. Particularly, network is compounded of different technologies that coexist each other, where some technologies could totally disable transmission of data of equipment that use other technologies. Thus, we need a solution that efficiently handles resources in space, frequency, and device dimensions. Efficient solution that can be used is semantic coordination in such networks [116,148].

The nodes in the system can communicate and share knowledge of their perspective of the spectrum utilization in the network. In [148] authors proposed to model the spectrum usage coordination as an interactive process between a number of distributed communicating agents, where agents share their specific information and knowledge. The information includes the current spectrum usage state, spatial coordinates of the device, available communication protocols, usage policy, spectrum sensing capabilities of the device, spectrum needs, etc. Approach for such coordination presented in [148] is based on semantic technologies, and harmonize communication between heterogeneous agents with potentially different capabilities with a minimal common compliance. The core knowledge is represented by ontologies whose representation and usage is specified in a standardized way. The approach is used as dynamic spectrum coordination algorithms used for coordination among different wireless technologies in 5G networking [116,148]. This semantic technologies based approach can be used for wide diapason of problems within 5G heterogeneous networks, such as network states predictions, network analysis, minimizing traffic load, content distribution coordination etc. This approach could be used in combination with caching techniques in order to improve content distribution in 5G, but further research should be done in this area.

1.2.4. Mobile CDNs and the Cloud

Mobile computing (MC) has created enormous demand for online experience, that OSN-aware CDNs are required to satisfy. Almost ubiquitous Wi-Fi coverage and rapid extension of mobile-broadband provide undisrupted connectivity for mobile devices, whereas devices that hop seamlessly from WiFi to cellular networks, and technologies such as 5G, will be optimised for uses that put a premium on continuous connectivity regardless of the user location [85]. Mobile-specific optimizations for applications along with drastically simplified and more intuitive use of devices (e.g. with multi-touch interactions instead of physical keyboards) contribute to mobile applications becoming the premium mode of accessing the Internet, at least in the US [82].

Cellular networks have become the main way citizens connect to the Internet worldwide, especially in developing countries. Thanks to the development of mobile devices and their networking capacities, as well as the arrival of fast and reliable networks such as 5G, a high quality connectivity is ensured everywhere and any time. The irruption of new paradigms, such as IoT, has increased the number of connected devices (sensors, actuators, etc.) which requires infrastructures that provide higher throughput networking, especially in use cases where high definition videos are involved and even new scenarios are yet to emerge.

Mobile Computing entails the processing and transmission of data over a medium, that does not constraint the human-medium interaction to a specific location or a fixed physical link. Fig. 3 depicts a general overview of the MC paradigm in its current form. It is

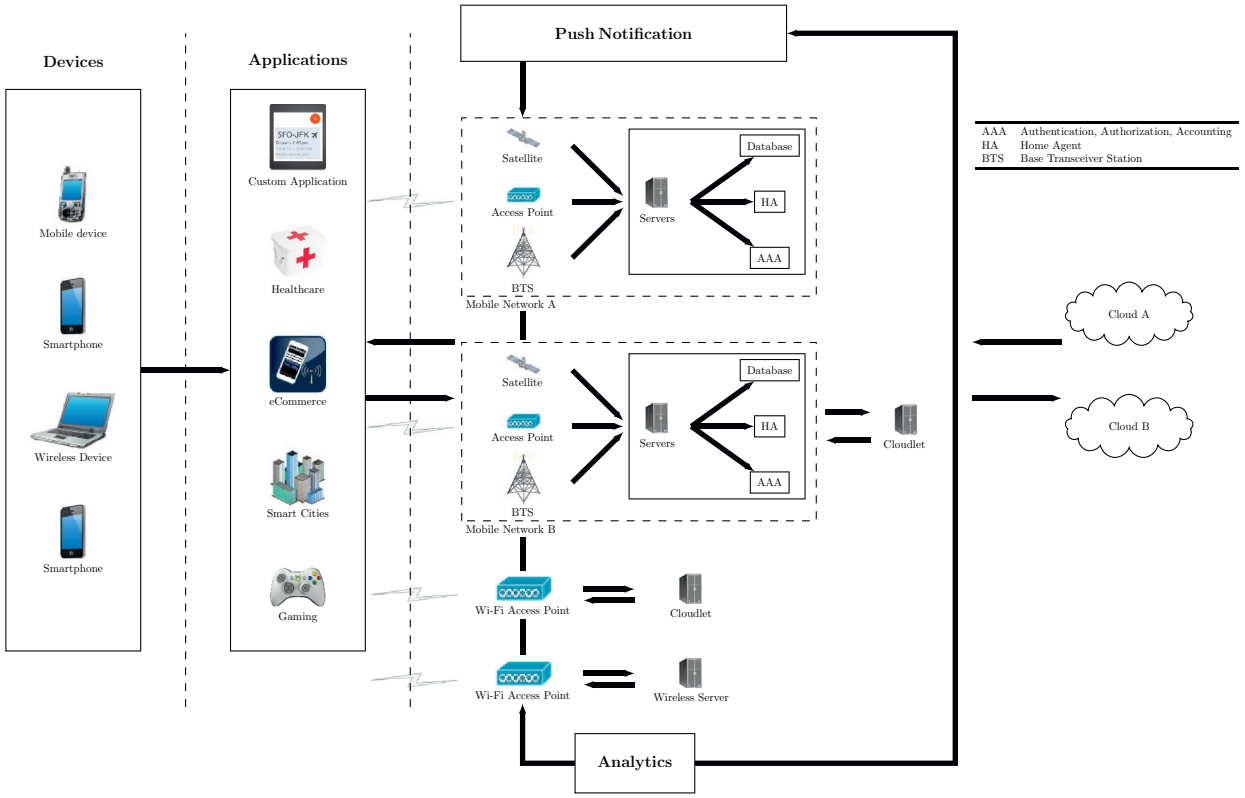

Figure 3. Social data streams over mobile computing.

the present decade that signifies the proliferation of MC around the world, although handheld devices have been widely used for around two decades in the form of Personal Digital Assistants (PDAs) and early smartphones. Almost ubiquitous Wi-Fi coverage and rapid extension of mobile-broadband (around 78 active subscriptions per 100 inhabitants in Europe and America) provide undisrupted connectivity for mobile devices, whereas 97% of the world's population is reported to own a cellular subscription in 2015 [85]. Mobile-specific optimizations for applications along with drastically simplified and more intuitive use of devices (e.g. with multi-touch interactions instead of physical keyboards) contribute to mobile applications becoming the premium mode of accessing the Internet, at least in the US [82]. Moreover, the MC paradigm is nowadays further combined with other predominant technology schemes leading to the paradigms of Mobile Cloud Computing [12], Mobile Edge Computing [111], Anticipatory Mobile Computing [120], etc. Today's mobile devices include smartphones, wearables, carputers, tablet PCs, and e-readers. They are not considered as mere communication devices, as they are in their majority equipped with sensors that can monitor a user's location, activity and social context. Thus, they foster the collection of Big Data by allowing the recording and extension of the human senses [102].

Mobile social networking involves the interactions between users with similar interests or objectives through their mobile devices within virtual social networks [21]. Recommendation of interesting groups based on common geo-social patterns, display of geo-tagged multimedia content associated to nearby places, as well as automatic exchange of data among mobile devices by inferring trust from social relationships are among the possible mobile social applications benefiting from real-time location and place information.

1. *Industrial applications*: Maintenance, service, optimization of distributed plant operations is achieved through several distributed control points, so that risk is reduced and the reliability of massive industrial systems is improved [122].
2. *Automotive applications*: Automotive applications capture data from sensors embedded in the road that cooperate with car-based sensors. They aim at weather adaptive lighting in street lights, monitoring of parking spaces availability, promotion of hands-free driving, as well as accident avoidance through warning messages and diversions according to climate conditions and traffic congestion. Applications can promote massive vehicle data recording (stolen vehicle recovery, automatic crash notification, etc.) [149].
3. *Retail applications*: Retail applications include, among many others, the monitoring of storage conditions along the supply chain, the automation of restocking process, as well as advising according to customer habits and preferences.
4. *Healthcare & telemedicine applications*: Physical condition monitoring for patients and the elderly, control of conditions inside freezers storing vaccines, medicines and organic elements, as well as more convenient access for people in remote locations with usage of telemedicine stations [81].

5. *Building management applications*: Video surveillance, monitoring of energy usage and building security, optimization of space in conference rooms and workdesks [149].
6. *Energy applications*: Applications that utilize assets, optimize processes and reduce risks in the energy supply chain. Energy consumption monitoring and management ([140,160]), monitoring and optimization of performance in solar energy plants [152].
7. *Smart homes & cities applications*: Monitoring of vibrations and material conditions in buildings, bridges and historical monuments, urban noise monitoring, measuring of eletromagnetic fields, monitoring of vehicles and pedestrian numbers to optimize driving and walking routes, waste management [70].
8. *Embedded mobile applications*: Applications for recommendation of interesting groups based on common geo-social patterns, infotainment, and automatic exchange of data among mobile devices by inferring trust from social relationships. Applications that continuously monitor the user's physical activity, social interactions and sleeping pattern, and suggest a healthier lifestyle.
9. *Technology applications*: Hardware manufacture, among many others, is improved by applications measuring performance and predicting maintenance needs of the hardware production chain.

Roadmap: Our survey is organized as follows: Section 1 presents shortly the characteristics of social data streams and introduces a taxonomy for challenges of distribution over them. Section 2 discusses existent surveys concerning modelling, simulation and performance evaluation in the examined bibliographical field. The association of context-aware computing with social networks is given in Section 3. Infrastructure efficiency of deployed data centres for the distribution of social content is analysed in Section 4 in terms of software solutions, as well as data centre scheduling frameworks. Section 5 presets a categorization of most predominant models for the depiction of the information diffusion process in a social network. Section 6 discusses various architectures, systems and techniques for efficient content distribution based on social data streams, along with diverse studies that corroborate them as well as the way 5G network infrastructure affects the social data streams. Section 7 concludes and finally gives the outline of future research directions.

2. Related work

In a manner that resembles the utilization of social data streams Anjum et al. [6] review the deployment of peer-assisted content delivery solutions. They present challenges caused due to heterogeneity in user access patterns and the variety of contextual information, such as interests and incentives of Internet Service Providers, End-Users and Content Providers. Furthermore, Perera et al. [122] survey context awareness from an IoT perspective, as they recognize that the emerging IoT is expected to enable expansion of conventional content delivery systems to a broader network of connected devices. They systematize the collection, modelling, reasoning, and distribution of context in relation to sensor data in a work that resembles the social data harvesting in terms of volume, variety and velocity. Their survey addresses a broad range of methods, models, systems, applications, and middleware solutions related to context awareness in the realm of IoT, that could be potentially applicable to social data streams, too.

In [96] Kilanioti et al. study various experiments on a modified content delivery simulation framework and compare miscellaneous policies for dynamic content delivery based on analysis of social data streams. The incorporation of an OSN-aware dynamic mechanism becomes indispensable for content delivery services, since (i)significantly large proportion of Internet traffic results from -easily produced via online media services and transmitted over OSNs- bandwidth-intensive multimedia content and (ii) multimedia content providers, such as YouTube, often rely on ubiquitous content distribution infrastructures. The policies presented take patterns of user activity over OSNs and exploit geo-social properties of users participating in extensive retransmissions of items over OSNs. The authors proceed to incorporate diverse caching schemes of the underlying infrastructure, miscellaneous policies for the handling of OSN data and various approaches that take into account the most efficient timing for content placement. The simulation framework introduced in [92] serves in this study as the basis of further parameterized content delivery experimentation that exploits information transmission over OSNs and decreases replication costs by selectively copying items to locations where items are bound to be consumed.

Downloads of large size multimedia contents are explored through several studies together with techniques that try to reduce doubled content transmissions using intelligent caching strategies in mobile networking [5,11,87]. The main idea is redistribution of mobile multimedia traffic in order to eliminate duplicated downloads of popular contents. Intelligent caching strategies would enable access to popular contents from caches of near node of mobile network operator. Those strategies allow content providers to reduce access delays to the requested content. Many caching algorithms for content distribution already exist [5]. Efficient caching strategy could enhance the energy efficiency of 5G networks, thus the cooperative caching architecture is presented in [87]. This strategy addressed the increasing demand for mobile multimedia and data services in energy efficiency in emerging 5G systems using content caching and distribution.

We are not aware of surveys in the bibliography suggesting an holistic approach for the utilization of social data streams towards facilitation of content distribution decisions and social analysis tasks. The diverse parameters we review in this work (modelling, simulation, performance evaluation) take into account low-level decisions and high-level considerations, including energy efficiency of employed data centres, in-memory keeping solutions and various network approaches for time-critical applications. We review combined aspects such as optimal route selection, data redundancy, data localization and data centre optimizations.

3. Social networks and context-aware computing

A social network is essentially a network of social bindings between people. Computer-Supported Cooperative Work (CSCW) has

contributed much in offering advanced collaborative systems for leveraging human connections and improving human interactions in workspace environments, but these systems mostly focus on business-driven interactions where connections among people tend to be formal and structured [27]. Recently however, social and computing disciplines focused specifically on the design of social-networking services, i.e. applications that support human social interactions and can be more informal.

On another dimension, the advancement of wireless networks, as well as mobile, context-aware and ubiquitous computing, enabled the improvement of social-networking services by enabling social encounters between proximate users with common interests in an anywhere and anytime fashion, as in Ubiquitous Computing systems [27]. There has been a shift thus of the application focus from virtual to physical social spaces using ubiquitous technologies [27]. This shift introduces a great number of possibilities, however it also introduces a number of challenges that are related to ubiquitous computing. While social-network systems for ubiquitous computing environments are an emerging trend in social computing, due to the fact that ubiquitous-computing environments are more dynamic and heterogeneous than Internet based environments, appropriate solutions and design guidelines are required to facilitate their ubiquitous aspect.

The term Ubiquitous Computing, first introduced in the nineties, refers to the shifting of the computing paradigm from the desktop Personal Computer (PC) to a more distributed and embedded form of computing [155]. Together with Pervasive Computing (for many these terms are synonymous), Ubiquitous Computing introduced the concept of “anywhere, anytime computing”, allowing users to interact with computers embedded in every-day objects in an “anywhere and anytime” manner. Moreover, Ubiquitous Computing specifies that the interaction of users with such devices must be straightforward in the degree that the user would not even notice such an interaction. In other words, in order for ubiquitous and pervasiveness to be achieved, computers must disappear from the front-end, be embedded to common objects that humans use daily and provide computational and informational services without expecting from users to explicitly and consciously interact with them.

Want and Pering [155], categorize the challenges in Ubiquitous Computing to: (i) power management issues - how mobile devices deal with processing power and storage space and the kind of wireless technology to use in every given situation, (ii) limitations in connecting devices – how are all these small devices going to be connected and managed, (iii) user interface issues – since Ubiquitous Computing demands for many different small-scale devices of various types of interfaces and displays of various sizes, the challenge lies in developing user friendly interfaces, (iv) issues related to Location Aware Computing. Henricksen et al. [76] add to the above list the challenge of managing heterogeneous devices of different hardware and software specifications, such as sensors and actuators, embedded devices in objects such as shoes, home and office appliances such as videos, mobile devices and traditional desktop computers, in order for these devices to interact seamlessly. Another challenge they mention has to do with maintaining network connections while devices move between networks of different nature and characteristics. In ubiquitous environments, people tend to use many devices simultaneously, therefore there is a need for these devices to communicate and exchange data. Another challenge Satyanarayanan [133] notes is tracking user intentions. This is important in Pervasive Computing in order for the system to understand what system actions could help the user and not hinder him/her.

Regarding context-awareness, an important challenge is to build context-aware systems that detect and manipulate the context in a human-like manner, i.e. making decisions proactively based on the context and provoke actions based on those decisions that assist the user through his/her task; the aforementioned should be done without any user participation or disturbance, except maybe in case of emergency. Another issue is how to obtain contextual information. Contextual information can be any information related to the user, the computing system, the environment of the user and any other relevant information regarding the interaction of the user and the system [43]. The user’s personal computing space can be used as the user’s context (any information regarding the user taken from her personal profile, calendars, to-do lists etc.), various types of context can be sensed in real time like location, people and objects nearby, while contextual parameters could also include the current emotional and physiological state of the user. Contextual challenges also include the way context is represented (ontologies can be used or other context modeling techniques), the way this information is to be combined with the system information, as well as how frequently should context information be considered. Hinze and Buchanan [74] differentiate the static context (e.g. user’s profile information) from the fluent context (dynamic, real-time context, e.g. time) and propose that a context model should be defined for each important entity, such as the user, the locations, etc. The authors mention as challenges the capturing of the context (whether it should it be done automatically at particular times or manually by the user) and the process of storing the context (whether it should be stored on the client, on the server or both). On the process of accessing contextual information, Hinze and Buchanan propose that context-awareness can help in reducing the amount of data to be accessed real time, by pre-retrieving any relevant pre-known data, e.g. the static context [74]. This increases efficiency.

Another challenge in developing ubiquitous systems is user modeling. User modeling in ubiquitous environments is challenging since a user often changes roles according to the context and the current environment he acts into; the big challenge is how to capture these changes and how to react on them [74].

Location, as an important contextual parameter plays an important role in context-aware systems and ubiquitous systems. The type of location sensing technology to be used is one issue, privacy is another - should user privacy be sacrificed for location awareness and to what extent - while a third issue is the semantic (and contextual) representation of the location in order to utilize more contextual parameters than just the location itself. For example, by semantically representing locations, one can attach to them various information resources such as a webpage, a user profile, various objects with semantic representation etc. Schilit e al. [139] proposed the movement from the simplified concept of location to more contextually rich notions of place where people and activities should also be considered. Possible problems towards this concept include the difficult management of large scale positioning data, privacy concerns regarding location-awareness and the challenge of how to associate information objects, such as a web page, with a real-world location. Privacy issues regarding location-awareness are related to human psychology: users often consider privacy issues when their location is to be known by a system, but at the same time they provide private information such as credit card numbers

and addresses to online systems without hesitation. This happens because in the first case they simply do not see the benefit of providing their location to be used by a simple application (e.g. finding friends in the proximity), while at the latter case they clearly see the benefit of buying goods online. The authors also argue that the centralized nature of the most location tracking applications (having a central server on which all user personal data are stored) discourages users from providing any personalized information, because centralized data can be accessed by anyone, not only illegally (e.g. hackers) but also the government, corporations with interest in user data (e.g. advertisers) etc. As a solution, a proposal is a decentralized schema where any personal data is stored and calculated on the client side, i.e. the user's device. An example of such a technology is the well known Global Positioning System (GPS): the client device uses satellite links to calculate locally the user's current position.

Context-Awareness and Adaptation related challenges and issues include

1. **Modelling the context:** which method is more appropriate to use
2. **Observing the context:** Automatically or manually
3. **Context sensing:** how are contextual parameters retrieved (sensors, user profiles etc.). In retrieving context data from various sources (e.g. sensors), how are inconsistencies between these data resolved
4. **Accuracy of contextual information** should be well known during the design of ubiquitous systems
5. **Storing the context:** on server (privacy issues), on client or on both
6. **Accessing the context**
7. **Using the context**
8. **How are the user and the environment connected and interact**
9. **How will the application modify its behaviour (be adapted) based on the context**
10. **Systems should be more context-aware than just the location.** A place is more than a location (also a Location related challenge)
11. **Devices should not operate based only on their own context, but based on the context of the whole system**
12. **Contextual information should be used to reduce the amount of input that is needed from users** (also a Human-Computer Interaction related challenge)
13. **How to capture changes in the user's role** – has to do with capturing the current context (i.e. the environment and the various circumstances) and user modelling (what possible role could a person play according to context)
14. **Context should be processed and various components should adapt to it without interfering with user's task** – no user explicit interaction should be necessary
15. **Adaptation in ubiquitous environments:** may need to adopt various devices separately and at the same time, while the user maintains a consistent view for the system/application

Context-aware computing has evolved over time from desktop applications, web applications,mobile computing, pervasive/ubiquitous computing to IoT over the last decade [122]. Context-aware computing became more popular with the introduction of the term 'ubiquitous computing' by Mark Weiser, while the term 'context-aware' was first used by Schilit and Theimer [142] in 1994. Context-aware computing has proven to be successful in understanding sensor data. Advances in sensor technology have evolved sensors in becoming more powerful, cheaper and smaller in size, which is the main reason huge numbers of sensors are deployed in the environment. This number is expected to grow over the next decade [137], generating ultimately big data [122,123].

In settings where social communities become mobile, i.e. users not only interact, meet and communicate via social networks, but are mobile as well (move into the environment, interact with others, etc.), the concept of group awareness is met [37,114,163] where context related to the group is exploited to enable ubiquitous applications and services to function and serve people's concerns and needs in a pervasive manner. There is a need, thus, for formulating dynamic communities aiming to facilitate people in performing common tasks. It is often the case that such dynamic communities are resolved after the current goals have been achieved [114]. It is evident, thus, that the context within which such dynamic communities are created, act, achieve goals and are then resolved is important, and that, through this context, we can understand the groups' interests and, thus, personalize the applications and services offered [114].

Through a bibliography study of mobile social network applications and platforms in [114] it is noted that the context features that these applications and platforms use can be summarized as follows: Location, Interest, Time, Personal, Activity and Social Interaction. Here, context is "any information that can be used to characterize the situation of an entity" [48] and social context is "the information relevant to the characterization of a situation that influences the interactions of one user with one or more other users" [156]. Moreover, in [114] a context-aware Mobile Social Network model is proposed aiming to facilitate the creation of dynamic social networks based on a combination of multiple contexts, including location, users" profile, domain specific data and OSN data, along with services for fostering the interaction among users.

4. Infrastructure efficiency

4.1. Software solutions for infrastructure efficiency

Various efforts have been made in order to optimize resource and energy efficiency for data centres that support social data interaction and analysis. The industry has made several efforts to address challenges [23,108,146] such as: (a) data growth, (b) isolation, (c) real-time interactions, (d) data centre federation and (e) cdn-load-balancing at data centre level, but usually lacks from

Table 1
Related work summary.

Ref [56]	Title: Performance evaluation of a green scheduling algorithm for energy savings in cloud computing		Savings ~ 45%
	Category:	Workload consolidation and power off policies (power off policy based on a neural network predictor)	
	Evaluation:	[8–512] nodes cluster simulation	
	Workload:	End user homogeneous requests that follow a day/night pattern	
Ref [107]	Title: Energy efficient utilization of resources in cloud computing systems		Savings [5–30]%
	Category:	Workload consolidation and power off policies (energy-aware task consolidation heuristic based on different cost functions)	
	Evaluation:	Simulation of a not stated size cluster	
	Workload:	Synthetic workload in terms of number of tasks, inter arrival time and resource usage	
Ref [125]	Title: Saving energy in data centre infrastructures		Savings [20–70]%
	Category:	Workload consolidation and shut-down techniques (safety margin power-off policy)	
	Evaluation:	100 and 5000 nodes cluster simulation	
	Workload:	Synthetic workload that follows a day/night pattern	
Ref [17]	Title: Energy efficient resource management in virtualized cloud data centres		Savings ~ 80%
	Category:	VM consolidation and migration	
	Evaluation:	100 nodes cluster simulation using CloudSim	
	Workload:	Synthetic workload that simulates services that fulfil the capacity of the cluster	
Ref [86]	Title: Dynamic energy-aware scheduling for parallel task-based application in cloud computing		Savings [20–30]%
	Category:	Energy-aware scheduling algorithms (polynomial-time and multi-objective scheduling algorithm for DAG jobs)	
	Evaluation:	Experimentation on a 64 nodes cluster	
	Workload:	Synthetic directed acyclic graph-based workload	

focusing on cost effectiveness. One of the main costs of these infrastructures is the energy consumption. In this context, an energy-efficient data-centre is measured as the computing and storage output produced from the energy consumed. The general indicator of a data centre efficiency is the so called Power Usage Effectiveness (PUE) [1] which is the *Total facility energy*/*Total IT energy* that has a theoretical minimum of 1.

To this end, as previously shown in Fig. 2, research community and others have developed solutions that can be arranged in the following categories:

- **Workload Consolidation and Shut-down Techniques**,
- **VM Consolidation and Migration**,
- **Data Replication and Placement, and**
- **Energy-aware Scheduling Algorithms.**

The common goal of these solutions is to minimize the idleness of the nodes and suppress resources and some examples for each category are shown in Table 1.

For **Workload Consolidation and Shut-down Techniques**, in [107] authors describe two energy-aware task consolidation heuristics, trying to maximize resource utilization so energy waste is minimized. To this end, these algorithms estimate the total CPU time consumed by tasks and prevent a resource from executing only one task. In [86], Juarez et al. proposal consists of an algorithm that looks for a minimum in a multi-objective function, considering the energy-consumption and execution time and by mixing a resource allocation solution and heuristic rules. They also evaluate this approach by simulating Directed Acyclic Graph (DAG) based workloads.

Moreover, other techniques of energy conservation are proposed, such as **Virtual Machine (VM) Migration and Consolidation** [14,17,18,143]. In [17], authors present a resource management approach for virtualized cloud data centres enabling lower energy consumption of the resources by applying VM allocation and migration based on current CPU usage. In addition, this work was extended by catering for Service Level Agreement (SLA) restrictions in [14]. These policies were evaluated over a 100-node cluster. In [143] the proposed algorithm is based on a Bayesian Belief network focusing on allocating and migrating VMs. In [125], authors present an energy manager for data centres that relies on the day/night patterns present on most of the Internet data centres. Its approach is based on aggregation of traffic during low usage periods and turning off of idle nodes.

For **Data Replication and Placement** to improve energy proportionality, many solutions [3,91,106,145] have been proposed. In [3], authors describe a power-proportional distributed file system which stores copies of data on non-overlapping subsets of resources. Each subset contains only one replica for each file and lets the administrator decide about the number of datasets that are going to be turned on to serve incoming requests, controlling, thus, the balance between energy consumption and performance. In a similar way, in [145] authors propose the division of the cluster in non-overlapping data zones (one replica per zone), and enable the administrator to power off the desired number of zones. In [91], Kaushik et al. present a modification of Hadoop File System (HDFS)

that divides the cluster in two zones depending on data usage patterns: the *Hot Zone* containing the recent data that is probably going to be accessed sooner and the *Cold Zone*, that contains the files with low spatial or temporal popularity with low probability to be accessed. Once it is divided in these two zones, a power off policy is applied to the *Cold Zone*. Finally, in [106], Luo et al. describe a non-uniform replica placement strategy based on data popularity. This way the number of available parallel replicas of data can be increased or decreased depending on its popularity.

Finally, for **Energy-aware Scheduling Algorithms** we can find various approaches that could enable powering off idle nodes, such as [56,86,88,107]. In [56], authors propose a *green* scheduling algorithm based on neural networks. This approach tries to predict workload demand in order to apply power-off policy to idle servers that are going to be underused or not used at all. Experiments performed simulate a medium or small data centre that runs a homogeneous workload composed of tasks deployed to satisfy end-user interactions. In [88], energy-aware scheduling policies combined with Dynamic Voltage and Frequency Scaling (DVFS) are presented in order to scale virtual resources.

Moreover, in order to have a clearer picture of the works by its purpose, we also classified them into three main focus: (a) modelling, (b) simulation, (c) performance.

4.2. Data centre scheduling frameworks

The aforementioned energy-aware scheduling algorithms solutions for infrastructure efficiency depend and are directly related with the data centre scheduling frameworks. The models for execution and negotiation of tasks and resources of several kinds of applications (MapReduce, interactive data analytics, web search jobs, etc.) have been transferred from traditional infrastructures to large-scale data centres [20]. Thus, data centres are shared by applications of several organizations and multiple users [19].

It is important to notice that each of these applications may be different in terms of the requirements of resource usage, latency, security constraints, etc. and thereby they differ in size, arrival timing and duration. This fact results in a set of heterogeneous workloads that have to be scheduled and placed. This scheduling is a big challenge, since each application may need an approach to achieve better performance: (a) Scalability for high job-arrival rates (b) High and Predictable performance; and (c) High resource utilization;. In order to improve the overall data centre efficiency, the utilization has to be increased. Thus, the latest approaches on resource allocation are focused on deploying multiple and heterogeneous workloads on the same set of hardware resources. The scheduling problem becomes more difficult due to diverse data processing.

We present the main categories of the scheduling frameworks based on various approaches and their limitations summarized in Table 3.

First generation of cluster schedulers were based on a Monolithic centralized approach. These schedulers [83] are good candidates under low job-arrival rate conditions, such as MapReduce jobs which usually are long-running [50]. This happens since this kind of workloads accepts latencies of seconds or minutes [57]. Monolithic centralized schedulers are able to perform high-quality decisions [52,65,161] by examining the cluster as a whole, in order to determine the impact of the heterogeneity of the hardware over the performance and interferences in shared resources [68,113,117,136,157]. This strategy tends to result in a higher machine utilization [153], shorter execution times and improvements in load balancing. It also leads to a more predictable performance [53,162], and the reliability of the whole system is increased [138].

It is important to notice that there is an increasing interest in fast-response jobs. This fact produced an amplification of the partitioning of jobs, and thereby a higher number of fast and smaller tasks to be scheduled. This new scenario may exceed the capabilities of the previous approaches of monolithic and centralized schedulers. This new challenge was faced by a couple of centralized scheduling solutions that parallelize scheduling actions:

- Two-level scheduling approach (Mesos [78], YARN [151]) features a centralized coordinator that perform a pessimistic approach to block the whole cluster while a scheduler is performing a scheduling decision. This coordinator offers resources to the upper frameworks such as Hadoop and Message Parsing Interface (MPI). Finally, these frameworks perform another scheduling decision for assigning tasks to nodes. This approach can be considered suboptimal since the whole cluster state and tasks requirements are not available for the involved framework scheduler (Table 2).
- Shared-state scheduling approaches (Omega [138]) feature a centralized coordinator that manages concurrent schedulers that can operate simultaneously. Thus, these approaches are known for following an optimistic approach. In this model, each concurrent scheduler performs scheduling decisions based on a stale copy of the cluster state. Next, they can commit atomic transactions to the centralized node. In case of conflict, the local copy of the scheduler is updated and the scheduling process starts again.

The centralized nature of the cluster management operation results in a performance bottleneck at the centralized coordinators. Thus, this centralized approach is suboptimal for environments where thousands or even millions of latency-sensitive and user-facing tasks should be deployed per second [119].

On the other hand, distributed schedulers [54,119,124,127] present a better performance under these previously presented conditions. This approach uses simple and fast algorithms that examine smaller parts of the cluster state, achieving higher throughput and low latency task placements. This also implies that scheduling decisions are worse in general.

However, the kind of workloads executed by large data centres has been evolving. Current scenarios present workloads that are heterogeneous [35,132] and composed by two types of jobs: (a) 10% of long-running jobs consuming 85% of the data centre resources; and (b) 90% of short and latency-sensitive jobs consuming 15% [4,126,131,161] In this scenario, poor placement decisions can specially have an impact on the long-running jobs. So distributed schedulers may achieve even worse performance than

Table 2
Classification of approaches according to its purposes.

Reference	Modelling	Simulation	Performance
[88]			
[86]			
[107]			
[14]			
[18]			
[143]			
[125]			
[56]			
[3]			
[91]			
[106]			
[145]			

Table 3
Cluster scheduling approaches.

Frameworks	Strategy	Optimal environment	Near-optimal environments
Paragon [52]	Centralized	Low number of long-running	Mid and high number of jobs
Quasar [53]	Monolithic	and non-latency sensitive jobs	Mixed workloads
Borg [153]			
YARN [151]	Centralized	Mid number of diverse	
	Two-level	long-running, non	Latency-sensitive workloads
Mesos [78]		latency sensitive jobs	
Omega [138]	Centralized	Mid number of heterogeneous	High number of short and
	Shared-state	workloads	latency-sensitive jobs
Canary [124]		High number of short,	Long, resource-demanding
Tarcil [57]	Distributed	non resource-demanding,	jobs
Sparrow [119]		latency-sensitive jobs	Mixed workloads
Mercury [97]	Hybrid	Mixed workloads composed of	Homogeneous workloads
Hawk [47]		90% of short, latency-sensitive	Other workloads patterns
Eagle [46]		jobs and 10% of long-running,	Evolving patterns
		resource-demanding jobs	

centralized scheduling approaches.

Hybrid solutions [46,47,97] came to overcome these limitations. These solutions mix both strategies (centralized and distributed) delivering centralized high-quality scheduling decisions for long-running and resource-demanding jobs and at the same time near optimal quick scheduling actions for latency-sensitive and short jobs.

We can conclude that scheduling frameworks have evolved to overcome ever-changing workload demands and requirements. Once a new computing paradigm arises, a new fine-tuned scheduling approach is developed to meet its requirements. Thus, new scheduling frameworks are currently being developed to solve particular problems [72].

Beyond scheduling, and to increase server utilisation, other approaches have focused on right-sizing server deployments by analysing the workload needs to avoid overprovisioning of resources at planning time [28]. Another area is right-sizing infrastructure redundancy to application needs [34], while formulating the right Key Performance Indicators (KPIs) to capture whole system efficiency.

5. Content diffusion models for social data streams

This section describes the most predominant models for the depiction of the content/information diffusion process in a social network. Most of the existent algorithmic solutions for content distribution are built on them, thus the assumption that content circulation over social data streams is depicted by one of them is of crucial importance for the suggested solutions. The main algorithmic problems studied in the bibliography are related with the discovery of nodes that are most prone to diffuse content to the greatest extent, and the categorization of nodes according to their influence degree. The categorization of the models is depicted in Fig. 4. The models presented are the most recent in the bibliography and there are no prior recent models to the best of our

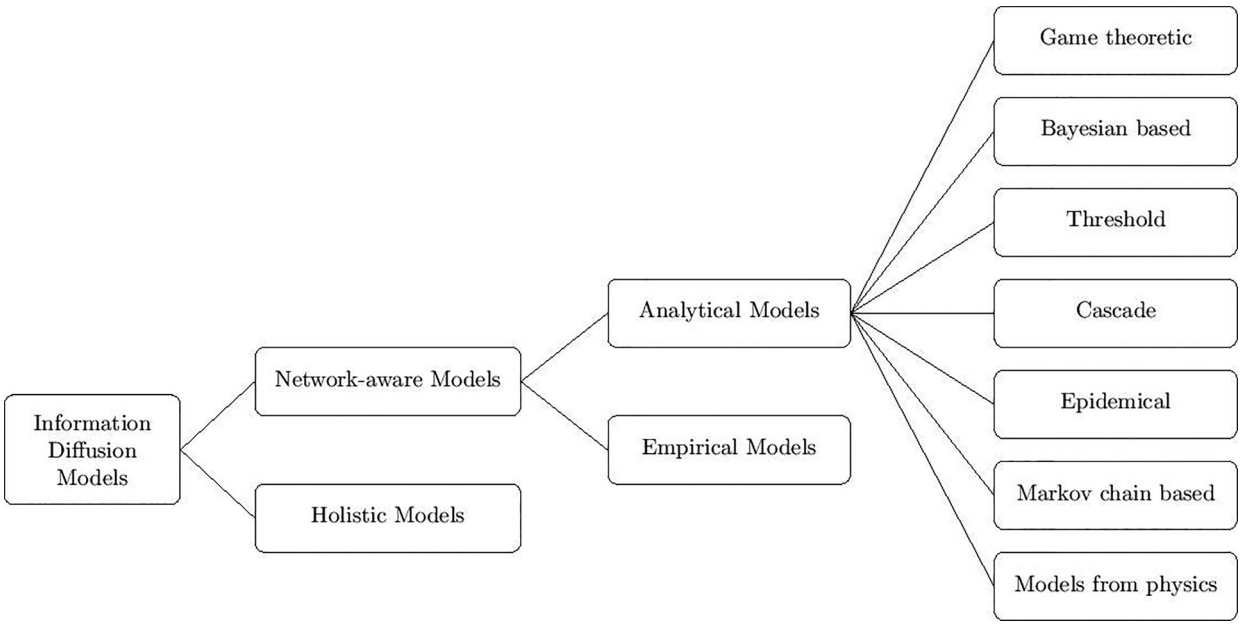

Figure 4. Content/information diffusion models.

knowledge. The first-level discrimination of models is based on whether they take the structure of the network into consideration (*network-aware*) or not (*holistic*). In other words the discrimination criterion is if they incorporate knowledge about underlying associations of the nodes (edges) or, to the contrary, follow an aggregate-level approach.

5.1. Holistic view models

Rogers' theory [130] is quantified by the Bass model [16]. The Bass model is based on the notion that "the probability of adopting by those who have not yet adopted is a linear function of those who had previously adopted" (F.Bass). It predicts the number of adopters $n(t) \in N$ of an innovation at time t (in the information diffusion scenario the number of retransmitters of an information piece):

$$n(t) = pM + (q - p)N(t) - q/M(\ N(t))^2 \tag{1}$$

where N(t) is the cumulative number of adopters by time t, M is the potential market (the ultimate number of adopters), $p \in [0, 1]$ is the coefficient of innovation (the external influences, expressing the individuals influenced by the mass media), and q is the coefficient of imitation (internal influence, expressing the individuals influenced by the early adopters). This approach, however, largely ignores the underlying network structure.

Models under the same concept of holistic view of the social behaviour make use of differential equations, and include, among others, the "multi step flow model" by Katz and Lazarsfeld [94], the Daley–Kendall rumours model [51], and also, more recent ones, such as, the Van den Bulte and Joshi model of influentials and imitators [150].

5.2. Network-aware models

These include completely novel models, but also variations of the afore-mentioned (holistic) models, such as the Nekovee variation [118] of the Daley-Kendall model, and are separated in the two following categories, based on whether they are mathematically formulated (*Analytical models*) and then applied or are the outcome of empirical methods, such as regression, regression trees etc. (*Empirical models*).

5.2.1. Analytical models

The first mathematical models based on nodes' thresholds for the depiction of information diffusion were developed by Schelling [135] and Granovetter [71]. There follows a categorization of the most predominant ones.

Game-theoretic models. In [95], Kleinberg proposes a simple networked coordination games model. We assume that there are two behaviours a node $v \in V$ in the graph $G = (V, E)$ can follow, A and B. The model is based on the notion that for each individual the benefits of adopting a new behaviour increase as more of its neighbours adopt the new behaviour. At discrete time steps each node updates its choice of A or B according to the behaviour of its neighbours. The objective of the nodes is to switch each time to the behaviour that reaps the maximum benefit for them. For the nodes *v* and *w* there is a motivation for behaviour matching, expressed in

the following way, where parameter q is a real number $0 < q < 1$:

– if v and w both choose behaviour A, they both receive a q payoff

– if v and w both choose behaviour B, they both receive a $1 - q$ payoff

– if v and w choose different behaviours, they both receive a 0 payoff

v's payoff for choosing A is qd_v^A and for choosing B is $(1 - q)d_v^B$. The overall payoff for v playing the game with its neighbours in G is the sum of the individual (pairwise) payoffs; q is actually the threshold expressing the fraction of adopting neighbours, since it easily results that v should adopt behaviour B if $d_v^B > qd_v$, and A if $d_v^B < qd_v$, where d_v is the degree of the node, d_v^A the number of its neighbours with behaviour A and d_v^B the number of its neighbours with behaviour B.

Initially there is a set S of nodes adopting behaviour B and $h_q(S)$ is the set of nodes adopting B after one round of updating with threshold q. $h_q^k(S)$ is the set of nodes adopting B after k successive rounds. A set S is *contagious* (with respect to h_q) if "a new behaviour originating at S eventually spreads to the full set of nodes" and the contagion threshold of a social network G is "the maximum q for which there exists a finite contagious set".

The technical issue of *progressive* or *non-progressive processes* (monotonous or non-monotonous as referred to later on in the present study) refers to the fact that when a node v following till then the behaviour A updates to behaviour B in time step t, it will be following B in all subsequent time steps. Although, intuitively, we would expect progressive processes to give finite contagious sets more easily (because of lack of early adopters setbacks that would hinder the cascade), Kleinberg points out that both the progressive and non-progressive models have the same contagion thresholds [112], which in both cases is at most 1/2 ("a behaviour can't spread very far if it requires a strict majority of your friends to adopt it") [112].

More game-theoretic models can be found in the work of Arthur [10], who proposes a simple cascade model of sequential decisions with positive externalities, manifested by a term that adds to the payoff of a decision. Namely in the scenario of two competing products, the latter become more valuable as more users use them (for a social media site or a smartphone, for example, it will acquire better third-party applications and support as its users grow). Also game-theoretic models are introduced by Banerjee [15] and Bikhchandani et al. [26], that are based on influence not due to positive externalities, but because of information conveyed from earlier decisions. The proposed game-theoretic models, however, have the drawback of not taking heterogeneity into consideration, in the notion that all nodes have the same threshold, and all their neighbours contribute the same in making a node change its behaviour.

Bayes-based models. Combining nodes' private information and their observations of earlier adoptions, in [58], Kleinberg and Easley present a Bayes based model to formulate information cascades, answering questions such as "What is the probability this is the best restaurant given the reviews I have read and the crowds I see there?".

$$Pr[A|B] = \frac{Pr[A]Pr[B|A]}{Pr[B]} \tag{2}$$

Three factors are taken into consideration:

- The states of the world;
- Payoffs; and
- Signals.

The first factor expresses whether an option is good or bad (if a new restaurant is a good or a bad choice). Supposing that the two options of the world are K (the option is a good idea) and B (the option is a bad idea), the world is placed in K with probability p and in B with probability $1 - p$ ($Pr[K] = p$, $Pr[B] = 1 - Pr[K] = 1 - p$). Payoffs for a node v are defined as follows: -If v rejects the option, the payoff is 0.

-If v adopts a good idea, it receives a positive $v_g > 0$ payoff.

-If v adopts a bad idea, it receives a negative $v_b > 0$ payoff.

-If v adopts without any prior knowledge, the payoff is 0.

The signals refer to private information each individual gets about the benefit or not of a decision: a high signal (H) suggests that adoption is a good idea, whereas a low signal (L) suggests that it is a bad idea. If accepting is indeed a good idea, then $Pr[H|K] = q > \frac{1}{2}$ and $Pr[H|K] = 1 - q < \frac{1}{2}$. In the restaurant example the private information could be a review that an individual reads about the first restaurant, with a high signal corresponding to a review comparing it favourably to restaurant B. If choosing the first restaurant is indeed good, there should be a higher number of such reviews, so $Pr[H|K] = q > \frac{1}{2}$.

Kleinberg and Easley [58] consider how individual decisions are made using (Eq. (5.2.1)) when they get a sequence of independently generated signals consisting of a number of high signals and a number of low signals, thus making interesting observations about situations where individuals can observe others' earlier decision, but do not have access to their knowledge.

The basic propagation models on which most generalizations for information diffusion are based are the Linear Threshold Model (LTM) [71,135,154] and the Independent Cascade Model (ICM) [69] with many proposed extensions (LTM: [93,154], ICM: [67,69,93]) and also a proposed unification [93].

Linear threshold model. Based on the assumption that some node can be either active (adopts a new idea / transmits a piece of information) or inactive and taking into account the monotonicity assumption, namely that nodes can turn from inactive to active with the pass of time but not the opposite, we can say that the LTM is based on the following notion: Each node v has a predefined

activation threshold $\theta_v \in [0, 1]$, which expresses how difficult it is for the node to be influenced when its neighbours are active ("the weighted fraction of the neighbours of node that must become active in order for node to become active"), and is influenced by each one of its neighbours w according to a weight b_{vw}, so that $\Sigma_{w \in \Gamma(v)} b_{vw} \leq 1$. The thresholds can be produced randomly with a uniform distribution, but some approaches investigate a uniform threshold for all the nodes of the network, e.g. Berger [22].The process takes place in discrete steps and the nodes that satisfy the constraint $\Sigma_{w \in \Gamma(v)} b_{vw} > \theta_v$ are gradually added as active to the initial set of nodes. It's worth mentioning that LTM can result as a modification of the networked coordinations game referred in the previous paragraph with the differentiation of payoffs for different pairs of nodes.

LTM expresses the idea that the influence of the neighbours of a node is additive, but when the rule of influence can not be expressed by a simple weighed sum, for example a node becomes active when one of its acquaintances and two of its co-workers do so, the arbitrary function g_v substitutes the weighed sum. In the *General Threshold Model* for time steps $t = 1, 2, 3...$ a node v becomes active if the set of active neighbours at t satisfy $g_v(X) > \theta_v$.

Independent cascade model. Under the ICM model [69], there is also a set of initially active nodes, the process takes place in discrete steps, but when node v becomes active, it has only one chance of activating each of its inactive neighbours w until the end of the process with a probability p_{vw} independent of the activations history and with an arbitrary order.

Exact evaluation of activation probabilities is exponential to the number of edges of the graph. Improving the performance of the works in [71,134], there are works studying the calculation of these probabilities such as Goyal et al. [66](based on a General Threshold Model with the assumption that each parent's influence is fixed), or Dickens et al. [55] (based on the ICM). In the latter, sampling from the twitter dataset is conducted in an efficient Markov-Chain Monte Carlo fashion using the Metropolis-Hastings algorithm [38] and the problem is tackled with two differentiations, one of which considering the past paths of data known (retweets for the twitter dataset) and one considering only the past path endpoints known (hashtags and urls) and joint probabilities are taken into consideration, reflecting also model uncertainty.

Epidemical models. In the case of epidemical models a single activated ("infected") node causes the change of state of a neighbour susceptible node, whereas in the afore-mentioned threshold and game-theoretic models a node has to interact with multiple neighbour nodes to evolve (complex contagion).

Epidemical models were introduced on the assumption that information would propagate like diseases. They constitute another category with an almost straightforward pairing with the ICM. The ICM captures the notion of contagion more directly, and also allows us to incorporate the idea that a node's receptiveness to influence does not depend on the past history of interactions with its neighbours.

Epidemical models variations include the simple branching processes model, where a node infects a number of nodes and the contagion proceeds in subsequent waves with a probability π. This model is characterized by the basic reproductive number of the disease $R_0 = k\pi$, where k is the number of new people somebody meets, which expresses the anticipated number of new cases of the disease that a single node will cause.

Extensions of the epidemical models are the SIR, SIS, and SIRS models: S stands for susceptible nodes, nodes that have not been infected yet and have no immunity to the contagion. I stands for infected nodes, nodes contagious to their susceptible neighbours, and R stands for recovered nodes, with the recovery considered as permanent in SIR and temporary in the case of SIRS [90]. The sequence of the letters in the acronyms of the models explains the flow of the epidemic. In SIR model nodes pass from the state of being susceptible to the state of being infected and then recover. In SIS model nodes are immediately susceptible once they have recovered (like in the case of common cold, recovery does not imply immunity that lasts for long). In the SIRS model recovered nodes free of infection may rejoin the susceptible nodes.

Markov chain models. Markov chains [45] are used to describe transitions from one state of a system to another in a finite set of possible states. Their memoryless nature (*Markov property*) has to do with the fact that the next state each time is independent of the preceding states. More formally: With a set of states $\{\xi_1, \xi_2, ., \xi_r\}$ the process moves successively from one state to another in so-called *steps*, and specifically from state ξ_i to state ξ_j with a probability p_{ij} (*transition probability*) independent of the previous states of the chains, or remains in the same state with a probability p_{ii}. A particular state is picked from Ξ as the initial state. Markov chains are usually depicted with a directed graph, where the edges' labels denote the transition probabilities.

Markov models are widely used for analysing the web navigation of users. PageRank [31] is based on a Markov model and is used for ranking of information in the World Wide Web. By assigning weights that denote the relative importance of an hyperlinked document in a set of documents, the likelihood that a person will reach a specific page through random clicks is, essentially, represented.

In [134], Song et al. use a Continuous-Time Markov Chain Model (CTMC), namely a Markov model that describes the transition among states after some time of stay in a particular state. This time is exponentially distributed and does not affect the transition probability to the next state. The information diffusion model is introduced on a network $G(V, w, \tau)$. G contains a set V of n nodes and E edges between nodes representing the information diffusion paths. w denotes the set of the edges' weights ("amount of information to flow from one node to another") and τ the set of the time delay on the information diffusion paths. Thus, the representation of the graph matches the CTMC in the notion that each node represents a state, each weight a transition probability and the delay is represented as the time-to-stay in each state.

Voter model. The basic voter model introduced by Clifford and Sudbury [42] and Holley and Liggett [79], is defined in an undirected

network and allows the spread of two opinions. In discrete time steps, a node adopts the opinion of a randomly chosen neighbour. For a node $v \in V$ in graph $G = (V, E)$, $\Gamma(v)$ is the set of neighbours of v in G and initially the nodes are arbitrarily endowed with a 0/1 state. At time step t each node adopts the opinion of one uniformly picked neighbour. With an initial assignment f_0: $V \rightarrow \{0, 1\}$ inductively we define

$$f_{t+1}(v) = \begin{cases} 1, with \quad probability \quad a \\ 0, with \quad probability \quad b \end{cases} \tag{3}$$

where $a = \frac{|\{u \in \Gamma(v) : f_t(u) = 1\}|}{|\Gamma(v)|}$ and $b = \frac{|\{u \in \Gamma(v) : f_t(u) = 0\}|}{|\Gamma(v)|}$.

Even-Dar and Shapira [60] argue that it is one of the most natural probabilistic models to capture the information diffusion in a social network. It is suitable for depicting the spread of a technological product, as it is proved that under this model consensus is reached with probability 1. Even-Dar and Shapira refer to the (almost) consensus of products such as Google as a search engine, YouTube as a video-sharing website etc.

Models from physics. Models from physics include the Ising model [84] serving for the description of magnetic systems, and bootstrap percolation [7] serving for the description of magnetic systems, neuronal activity, glassy dynamics, etc.

The Ising model [84] was first proposed in statistical physics and encompasses the notion of a ground state (in physics the state with the minimum energy), and that of the "self-optimizing" nature of the network.

Similarly to the basic voter model, there can be two competing "opinions", in favour of or against a subject, let's say depicted by a "+1" and a "-1", which in physics express the correspondence of an atom forming a network to a spin variable (can be considered as the basic unit of magnetization) state $\sigma_i = \pm 1$. The total energy of the system under this model (Hamiltonian) is defined as:

$$H = H(\sigma) = -\sum_{<i,j>} E\sigma_i\sigma_j - \sum_i J\sigma_i \tag{4}$$

for each configuration $\sigma = (\sigma_1, ..., \sigma_N)$, with the parameter J associated with an "external magnetic field" and E with the "nearest-neighbours interaction", N the number of the atoms. The ground state is the lowest energy configuration s_g (in physics the zero temperature configuration), so that $s_g \in argmin_s H(s)$. In a social network can be seen as the state with the most likely opinion, minimizing conflicts among its members (atoms).

In the standard bootstrap percolation process [7] a node is initially either active with a given probability f or inactive. It becomes active if k ($k = 2, 3, ...$) of its nearest neighbours are active. In that notion it resembles the k-core problem of random graphs [105], where k-core is the maximal subgraph within which all vertices have at least k neighbours, but whereas bootstrap percolation starts from a subset of seed vertices according to the above-mentioned activation rule, the k-core of the network can be found by a subsequent pruning of vertices which have less than k neighbours.

5.2.2. Empirical models

Before the advent of machine-readable traces, the potential of networks in the transmission of information and messages was stated already by Milgram in his renowned experiment [110] or Christakis [64], who suggested in a study of 12,000 participants that risks, such as the risk of becoming obese or benefits, such as stopping of smoking, are propagated through social ties. However, it is large scale and time-resolved machine-readable traces that, through the step-by-step track of interactions in OSNs (although not compulsorily easily accessible/ collectible), have driven to the formulation of a plethora of empirical models.

Some generic observations concerning the empirical models are the following. Many of them lack insight of information content, unlike works, such as that of Huberman et al. [2], who formulate a model taking into consideration solely the features of an information item (a news item in Twitter). Sometimes the discovered patterns in empirical models are at odds with the predictions based on theoretical (analytical) models. For example, in unison with the epidemical model, Leskovec et al. in [100] claim that cascades (depicting the blogosphere information diffusion) are mostly tree-like. More specifically, they notice that the number of edges in the cascade increases almost linearly with the number of nodes, suggesting that the average degree in the cascade remains constant as the cascade grows (a trees property). Moreover, Leskovec et al. claim that these trees are balanced, as they notice that the cascade diameter increases logarithmically with the size of the cascade. In contradiction to the above, the trees derived from the chain-letter diffusion model of Liben–Nowell and Kleinberg in [103] are inconsistent with the epidemic model, as they are very narrow and deep, with the majority of their nodes having one child and a median distance from their root to the their leaves being of hundreds steps.

Precisely, in [103] the spread of a chain-letter is represented by a tree. Copies of the chain-letter represent paths through the tree, the root represents the originator and the leaves represent the recipients of a message (w is a child of v if w appends its name in the copy of the letter directly below v). In order to produce trees with the characteristics mentioned in the previous paragraph, the probabilistic model suggested (i) incorporates asynchrony: after receiving a message, each recipient waits for a time t before acting on it, and if it receives more copies of the item in this time interval, it acts upon only one of them, and (ii) encompasses a back-rate β, as a node can either forward the message to its neighbours with probability $1 - \beta$ or group-reply to his corecipients with a probability β.

In [25], Bakshy et al. attempt to model the information diffusion in Twitter with the use of regression trees. Twitter is convenient for information diffusion modeling, since it is explicitly diffusion-oriented: users subscribe to the content of other users. The retweet feature, moreover, helps in the acknowledgement (though does not guarantee it) of reposts. Seeders are users posting original (not retweeted) content and reposting instead of the conventional retweeting (RT @username) is taken into account. Influence is

measured in terms of the size of the whole diffusion tree created, and not just the plain number of explicit retweets. The three different cases studied ascribe the influence to the first one having posted a link, the most recent one or follow a hybrid approach.

The predictors used include, for the seed users: the number of followers, number of friends, number of tweets and date of joining, and regarding the past influence of seed users: the average, minimum and maximum total influence and average, minimum and maximum local influence (local refers to the average number of reposts by a user's immediate friends over a period of one month and total to the average total cascade size over that period).

Bakshy et al. [25] come to the conclusion that although large cascades have in their majority previously successful individuals with many followers as initiators, individuals with these characteristics are not necessarily bound to start a large cascade. Thus, because of the fact that estimations cannot be made at an individual level, marketers should rely on the average performance. By studying the return on investment, on the whole, with a cost function of the number of followers per individual i: $c_i = ac_f + f_i c_f$, where a is acquisition cost c_f cost per follower and f_i is the number of followers, they conclude that relatively ordinary users of average influence and connectivity are most cost-efficient.

Content-related features are, also, according to Bakshy et al. not expected to discriminate initiators of large cascades from non-successful ones, due to the large number of non-successes. In order to take content into account the regression analysis is repeated encompassing the following features: rated interestingness, perceived interestingness to an average person, rated positive feeling, willingness to share via email, IM, Twitter, Facebook or Digg, some indicator variables for type of URL, and some indicator variables for category of content.

Moreover, Lerman et al. [98] claim that exploiting the proximity of users in the social graph can serve as an adding-value factor for the prediction of information diffusion. They discriminate proximity as coming from conservative or non-conservative processes (denoting that the amount of spread information in the network remains or not constant, respectively). For the case the underlying network is not fully known [115], Najar et al. focus on predicting the final activation state of the network when an initial activation is given. They find the correspondence between the initial and final states of the network without considering the intermediate states. Their work is based on the analogy between predictive and generative approaches for discrimination or regression problems (predictive models depicting a better performance, when the real data distribution can't be captured).

In [158], Yang and Leskovec use a time series model for modeling the global influence of a node through the whole network. For each node u, an influence function $I_u(l)$ is the number of mentions of an information l time units after the node u adopted the information (at t_u), and with $V(t)$ being the number of nodes that mention the information at time t, it applies:

$$V(t+1) = \sum_{u \in \in A(t)} I_u(t - t_u) \tag{5}$$

where A(t) are the nodes that got activated before t, $t_u \leq t$. For the modeling of the influence functions a non-parametric formulation followed allows greater accuracy and deviation, as no assumptions are made.

A study of the social news aggregator Digg [44] crawling data from the site, story, user and social network perspective, suggests the presence of previously unconsidered factors for the steering of information spread in OSNs. Doerr et al. suggest, that, beyond the bare OSN topology two factors matter: the temporal alignment between user activities (i.e. whether users are visiting the site in the same narrow time window) and a hidden logical layer of interaction patterns occurring in their majority outside the social graph.

In the direction of studying the information diffusion as social graphs evolve, Ren et al. [128] study the evolution steps for shortest paths between two nodes, (so that they can ascribe them to a disjoint path, a short-circuiting bridge or a new friend between them), and furthermore, metrics such as closeness centrality, and global metrics, like the graph diameter, across snapshots of gradually evolving graphs. To this end, they adopt an efficient algorithm and an efficient storage scheme.

Firstly, they cluster (in an incremental procedure not requiring all snapshots to be present in memory) successive graphs exploiting their many resemblances (daily snapshots). As $G_{\cup}$ and $G_{\cap}$ essentially "bound" the graphs in the cluster, with $G_{\cap}$ being the intersection (the largest common subgraph) of all snapshots in cluster C, and $G_{\cup}$ the union (the smallest common supergraph) of all snapshots in C, grouping of snapshots into clusters can be based in the idea of the graph edit similarity between these two graphs ($G_{\cup}$, $G_{\cap}$). The graph edit similarity to capture the similarity requirement of a cluster is defined as:

$$ges(G_a, G_b) = \frac{2|E(G_a \cap G_b)|}{|E(G_a)| + |E(G_b)|} \tag{6}$$

Secondly, they exploit the idea that, denoting the shortest-path between the vertices v and u, by $\tilde{P}_*(u, v)$ in a graph G_*, where *=1, 2,...,n, ∩,∪, the solution can easily be found by the intersection or union (two graphs) of graphs in the cluster, or be "fixed" using these two graphs, and they propose a "finding-verifying-fixing framework".

As far as the storage schemes variations are concerned, for a cluster of snapshots $C = G_1, ..., G_k$ the deltas $\Delta(G_i, G_{\cap})$, $\forall 1 \leq i \leq k$ consist a small fraction of the snapshot, and their size depends on the threshold value used for clusters' similarity. The penalty of decompression overheads needed is surpassed by savings in I/O. Variations of the storage schemes include the following:

$$SM1(C) = \{G_{\cap}, \Delta(G_{\cup}, G_{\cap}), \Delta(G_i, G_{\cap}) | 1 \leq i \leq k\} \tag{7}$$

$$SM2(C) = \{G_{\cap}, \Delta(G_{\cup}, G_{\cap}), \Delta(G_1, G_{\cap}), \mathcal{D}(G_i, G_{i-1}) | 2 \leq i \leq k\} \tag{8}$$

$$SM\_FVF(C) = \{\mathcal{D}(G_{\cap}, G_{p\cap}), \Delta(G_{\cup}, G_{\cap}), \Delta(G_1, G_{\cap}), \mathcal{D}(G_i, G_{i-1}) | 2 \leq i \leq k\} \tag{9}$$

In (7) the authors consider only the edge sets of $\Delta(G_i, G_{\cap})$ and $G_{\cap}$ to execute their algorithms on a snapshot G_i and the snapshots, G_i's, of the cluster need not be explicitly stored. For further compression of data of an evolving graph sequence similarity of

successive snapshots is exploited: In (8) $D(G_i, G_{i-1}) = (E_i^+, E_i^-)$, where $E_i^+ = E(G_i) - E(G_{i-1})$ and $E_i^- = E(G_{i-1}) - E(G_i)$ are the changes made to snapshot G_{i-1} to obtain the next snapshot Gi. Authors observe that the size of the set of edge changes $D(Gi, Gi - 1)$ is on average just 1/10 the size of $\Delta(G_i, G \cap)$. Hence, representing an EGS in terms of the D's is much more space efficient than in term of the Δ's. Further compression can be achieved by exploiting inter-cluster redundancy (9).

6. Distribution of social data streams

6.1. Content distribution for social data streams

6.1.1. Architectures

In [89], Jacobson et al. introduce Content Centric Networking (CCN), noting that network use has evolved to be dominated by content distribution and retrieval. CCN has no notion of host at its lowest level - a packet "address" names content, not location, while simultaneously preserving the design decisions that make TCP/IP simple, robust and scalable. Content is treated as a primitive, and with new approaches, Jacobson et al. simultaneously achieve scalability and performance.

To share resources within the context of a social network with the use of the cloud business model, Chard et al. in [36] propose the SocialCloud architecture. Users register in cloud services (computational capacity, photo storage etc.), and their friends can consume and provide these services through a Facebook application. The allocation of resources (trading or reciprocal use between friends) is conducted by an underlying market infrastructure, whereas the Social Cloud application passes a SLA to the service. The advertisement of the service, so that it can be included in the market is done with XML based metadata stored in Globus Monitoring and Discovery System (MDS).

An interesting approach [104] applicable to the realm of content delivery is based on an architecture which combines global learning and local caches with small population. It is shown that age-based thresholds can timely exploit time-varying popularities to improve caching performance. Moreover, the caching efficiency is maximized by a combination of global learning and clustering of access locations, accompanied by score mechanisms to help with practical issues at local caches. Practical considerations include, though, the size of the content that circulates over OSN and the long-tail effect, since the goal of the authors is first to learn a good estimate at the global point and then feed it back to the local caches in the form of content scores, thus, making the approach possibly prohibitive for OSN-aware content delivery.

6.1.2. Systems

In Buzztraq [144], Sastry et al. build a prototype system that takes advantage of the knowledge of the users' friends' location and number, to generate hints for the placement of replicas closer to future accesses. Comparing their strategy with location based placement, which instead uses the geographical location of recent users, they find substantial decrease of cost, when requests as part of cascades are more than random accesses of content. Furthermore, their system reacts faster when there is a new region shift, since it starts counting friends of previous users in a new region, even before a request comes from that region. The key concept of Buzztraq is to place replicas of items already posted by a user closer to the locations of friends, anticipating future requests. The intuition is that social cascades are rapidly spread through populations as social epidemics. The experimental results indicated that social cascade prediction can lower the cost of user access compared to simple location-based placement. Buzztrack is a simple system that only provides hints as to where to place objects. Other more complex constraints that the present work covers, such as server bandwidth and storage, are not taken into account. Moreover, social cascade is indirectly analysed because there has to be a third-party page where users connect to view the videos and have access to their social profile.

In the direction of distributing long-tailed content while lowering bandwidth costs and improving QoS, although without considering storage constraints, Traverso et al. in [147] exploit the time differences between sites and the access patterns that users follow. Rather than naively pushing UGC immediately, which may not be consumed and contribute unnecessarily to a traffic spike in the upload link, the system can follow a pull-based approach, where the first friend of a user in a Point of Presence (PoP) asks for the content. Moreover, rather than pushing content as soon as a user uploads, content can be pushed at the local time that is off-peak for the uplink and be downloaded in a subsequent time bin, also off-peak for the downlink. The larger the difference is between the content production bin and the bin in which the content is likely to be read, the better is the performance of the system.

In [141], Scellato et al. study how Twitter can be used to examine social cascades of UGC from YouTube and discover popular objects for replication. They improve the temporary caching policy by placing content after accounting for the distance between users. For the model CDN system constructed and tested, Scellato et al. used the Limelight network properties with 19 clusters of servers worldwide. To test the system, two different video weights were used: geosocial, in which node locality values are calculated from all the users that have posted a message about the item (even without being involved in a cascade), and geocascade, in which node locality values are calculated from the users participating in the item's social cascade. It was shown that the model improved performance against a no weight policy, with geocascade weight performing better.

6.1.3. Techniques

The introduction of concrete, unified metrics for the characterization of the extent of the social dissemination (local or global cascades phenomena) is an open issue. A systematic incorporation of this quantified knowledge into the existent underlying content delivery infrastructure would be salutary for proactive steps towards the improvement of user experience.

Furthermore, novel techniques aim to incorporate the information extracted from OSNs in the way that users share content and in how the content ultimately reaches the users. Some of these works use the information directly from OSNs, whereas others use such

information indirectly. The research goals vary: the decision for copying content, improvement of policy for temporary caching, etc.

Zhou et al. [164] leverage the connection between content exchange and geographic locality (using a Facebook dataset they identify significant geographic locality not only concerning the connections in the social graph, but also the exchange of content) and the observation that an important fraction of content is "created at the edge" (*is user-generated*), with a web based scheme for caching using the access patterns of friends. Content exchange is kept within the same Internet Service Provider (ISP) with a drop-in component, that can be deployed by existing web browsers and is independent of the type of content exchanged. Browsing users online are protected with *k*-anonymity, where *k* is the number of users connected to the same proxy and are able to view the content.

In [75], Hoque and Gupta propose a technique for putting with a logical addressing scheme together in the disk blocks containing data from friends. The large scale of OSNs and the predominant tail effect do not allow use of techniques such as those used in multimedia file systems or web servers, where items are globally popular, and, techniques keeping related blocks together tracking the access pattern of blocks, respectively. To this purpose, in [75] the social graph is divided into communities, and the organization of blocks in the disk is conducted with a greedy heuristic that finds a layout for the users within the communities and organizes the different communities on the disk by considering inter-community tie strength. The system is implemented on top of the Neo4j graph database as a layout manager.

Instead of optimizing the performance of UGC services exploiting spatial and temporal locality in access patterns, Huguenin et al., in [77], show on a large (more than 650,000 videos) YouTube dataset that content locality (induced by the related videos feature) and geographic locality are in fact correlated. More specifically, they show how the geographic view distribution of a video can be inferred to a large extent from that of its related videos, proposing a UGC storage system that proactively places videos close to the expected requests. Such an approach could be extended with the leverage of information from OSNs.

Kilanioti et al. [96] aspire to propose miscellaneous policies for dynamic OSN-aware content delivery over a content delivery simulation framework. The authors propose policies that take patterns of user activity over OSNs and exploit geo-social properties of users participating in social cascades, proceed to incorporate various caching schemes of the underlying infrastructure, different policies for the handling of OSN data and various approaches that take into account the efficient timing of prefetching. Given an efficient placement of surrogate servers with maximum performance and minimum infrastructure cost, they apply contextual features of the user as heuristics to find the best content diffusion placement, either in a global or in a local scale, i.e., which content will be copied in the surrogate servers and to what extent, not overlooking memory, time and computational cost. Moreover they study temporal aspects of diffusion, related to the most efficient timing of the content placement.

In terms of performance, Kilanioti et al. note a significant improvement over the respective improvement (39.43% only for the plain Social Prefetcher approach [92], upto 42.32% for selected caching mechanisms, compared to 30% in [147]) performing better than existent methods in pull-based methods employed by most CDNs, even though these methods additionally overlook storage issues of the distributed infrastructure.

Last but not least, of more concurrent cascades happening it would be interesting to know which of them will evolve as global and which of them will evolve as local, possibly making some associations with their content or context features. It is challenging to discover contextual associations among the topics, which are by nature implicit in the user-generated content exchanged over OSNs and spread via social cascades. In other words we would like to derive semantic relations. This way the identification of a popular topic can be conducted in a higher, more abstract level with the augmentation of a semantic annotation. While we can explicitly identify the topic of a single information disseminated through an OSN, it is not trivial to identify reliable and effective models for the adoption of topics as time evolves ([73], [101]) characterized with some useful emergent semantics. Therefore efficient semantic annotation can be seen as a solution for the challenge of characterization of the extent of the social dissemination.

6.2. Content distribution in 5G environments and technologies

Since content, especially video and multimedia content, which counts 80–90% of the total global traffic, is the main information item exchanged between different actors in the Internet with, current capacity of the wireless link and mobile core network cannot support such traffic.

5G brings high increasing ratio of mobility communications and strong orientation towards content-related services and applications for content delivery over wireless technology, high throughput, low data delivery latency, and high scalability enabling huge number of devices [13]. 5G opens new possibilities, increase radio link capacity and brings plenty of new trends such as [13,99]: heterogeneous networks (HetNets); new use cases based on connections and communications between device to device, massive Machine-Type Communications, and IoT; evolution of radio access technologies; cloudification through software defined networking (SDN) and network function virtualization (NFV) paradigms; flexible spectrum management; cell densification; etc. NFV and SDN capabilities in 5G systems are expected to enable network programmability. Cloudification of 5G through different SDN/NFV paradigms may significantly affect content delivery [99]. The 5G network includes programmable network control and allows the virtualization of all the RAN elements into virtual appliances by flexible NFV management, which enable content focused resources allocation. Since agile design of new network functions and their control becomes possible, network providers could extend network with new function that includes custom designed information. They can extend network with services that can offer to the online media service providers in order that they can choose the needed information and functional services.

Media delivery solutions designed for 5G can enable the collaboration between the network provider and the online media service provider by means of the edge cache. The network provider will keep the control of the network and give only the relevant information for the online media service provider, while the online media service provider will keep the control of the delivery process and decide whether the cache shall be used, what and how information or resources are cached [59].

New constituent technologies of 5G such as LTE-A, LTE-U, WiFi, ZigBee etc, SDN and NFV already starts to rapidly change networks and services and lead to changes to content delivery. It is predicted that mobile video will generate more than 69% of mobile data traffic by 2019 [39]. We witness the increase at 75% by 2021 which is much greater from 46% in 2016 on the share of smart devices and connections, while the amount of traffic offloaded from 4G was 63% at the end of 2016, and it will be 66% percent by 2021 [39]. By introducing 5G data speeds will be high and enable traffic to stay on the mobile network instead of being offloaded. In such case the offload percentage will be less than 50 percent and we will see higher offload rates with 5G network arrives. The main challenges in wireless or mobile environment that have impact on content delivery services are reflected in the limited spectrum and bandwidth in wireless, heterogeneous networks, wireless link characteristics that are dependent on location and time, radio congestion, handoff issues, etc.

Directions in future 5G developments is dependent on service providers, technology enablers and customers which are directly involved in decisions which use cases to pursue first as well what technology is needed for that usecases. Based on chosen use cases 5G standards development process is dependent, too. All these use cases and ongoing developments will directly affect content delivery mechanisms, models and systems architectures. For now, the main 5G use cases are reflected in [59]:

1. Gigabit broadband to home
2. Next generation mobile user experience
3. Future corporate networks
4. Digital industrial ecosystems
5. Infrastructure as a service

Gigabit broadband to the home use case is related to deliver streams rated from100Mbit/s to 1Gbit/s, which are needed to deliver television with higher resolution than 4K, virtual and augmented reality. Specific applications require special network configuration, for example in order to minimize latency in virtual reality applications. Next-gen mobile user experience is addressed to better service providing, which require operators to dynamically manage network and to use software defined networking and network function virtualization. Future corporate applications involve mining, autonomous driving, and robotics, which require to providers and customers work together on network defining and deploying. Digital industrial ecosystems include agriculture, smart cities and healthcare applications, which imply network configurations that every industry participant can benefit from. Next generation infrastructure as a service approach is for service providers that lack the resources to invest in nationwide 5G coverage.

7. Conclusions

Multimedia big data from entertainment and social media, medical images, consumer images, voice and video drives research and development of related technologies and applications and is steadily becoming a valuable source of information and insights. Multimedia content providers such as YouTube strive to efficiently deliver multimedia big data to a large amount of users over the Internet, with currently more than 300 hours of video content being uploaded to the site every minute. Traditionally, these content providers often rely on social data content distribution infrastructures. However, some measurement studies depict that a significantly large proportion of HTTP traffic results from bandwidth-intensive multimedia content circulating through OSNs. Consequently we can exploit the user activity extracted from OSNs to reduce the bandwidth usage. By incorporating patterns of information transmission over OSNs into a simulated content distribution infrastructure, the performance of content distribution mechanisms can be remarkably improved.

CDN services are increasingly being used to enable the delivery of bandwidth-demanding large media data to end-users of multimedia content providers and extend the capabilities of the Internet by deploying massively distributed infrastructures to accelerate content delivery. Next generation CDNs are being leveraged in an array of ways to overcome the challenges of providing a seamless customer experience across multiple devices with varying connectivity and corresponding to the call for enterprise application delivery. They have to go beyond efficient resource discovery and retrieval tasks of the established CDNs and support refined mechanisms for data placement, replication and distribution for a large variety of resource types and media formats. OSNs on the other hand create a potentially transformational change in user navigation and from this angle the rapid proliferation of OSNs sites is expected to reshape the architecture and design of CDNs. The challenges and opportunities highlighted in the interdisciplinary field of OSN-aware content delivery are bound to foster some interesting future developments, including innovative cache replacement strategies as a product of the systematic research of temporal, structural and geographical properties of social cascades.

Especially today that HTTP traffic ascribed to media circulating over OSNs has grown, an OSN-awareness mechanism over content distribution schemes has become essential. This mechanism aims to exploit patterns of social interactions of the users to reduce the load on the origin server, the traffic on the Internet, and ultimately improve the user experience. By addressing the issue of which content will be copied in the surrogate servers of a CDN, it ensures a near-optimal content diffusion placement. At the same time, it moderates the impact on bandwidth that the Big Data transmitted via OSNs has, offering scalable solutions to existing CDNs or OSNs providers. Furthermore, it paves the way for experimentation with variations on caching schemes, timing parameters of content delivery and context of the OSN and the media platform.

To conclude, cloud service providers have added CDN services in order to lower costs while increasing simplicity. CDNs, often operated as Software as a Service (SaaS) in cloud providers (Amazon CloudFront, Microsoft Azure CDN, etc.) aim at addressing the problem of smooth and transparent content delivery. A CDN actually drives cloud adoption through enhanced performance, scalability and cost reduction. With the limitation for both CDNs and cloud services being the geographic distance between a user asking

for content and the server where the content resides, cloud acceleration and CDN networks are both complementary to achieving a goal of delivering data in the fastest possible way. Although cloud mainly handles constantly changing and, thus, not easily cached dynamic content, utilization of OSN-aware CDNs in cloud computing is likely to have profound effects on large data download.

References

[1] V. Avelar, D. Azevedo, A. French, E.N. Power, Pue: AComprehensive Examination of the Metric, 49 (2012). White paper
[2] R.B.S. Asur, R. Bandari, B. Huberman, The Pulse of News in Social Media: Forecasting Popularity, 1202 Association for the Advancement of Artificial Intelligence, 2012.
[3] H. Amur, J. Cipar, V. Gupta, G.R. Ganger, M.A. Kozuch, K. Schwan, Robust and flexible power-proportional storage, Proceedings of the 1st ACM symposium on Cloud computing, ACM, 2010, pp. 217–228.
[4] G. Ananthanarayanan, A. Ghodsi, A. Wang, D. Borthakur, S. Kandula, S. Shenker, I. Stoica, Pacman: coordinated memory caching for parallel jobs, Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, USENIX Association, 2012. 20–20
[5] B. Assila, A. Kobbane, M. El Koutbi, A survey on caching in 5g mobile network, 2017.
[6] N. Anjum, D. Karamshuk, M. Shikh-Bahaei, N. Sastry, Survey on peer-assisted content delivery networks, Comput. Netw. (2017).
[7] M. Aizenman, J.L. Lebowitz, Metastability effects in Bootstrap Percolation, J. Phys. A 21 (19) (1999) 3801.
[8] Alexa, (http://alexa.com/topsites). [Online; accessed 20-Dec-2017].
[9] C. Anderson, Long Tail, The, Revised and Updated Edition: Why the Future of Business is Selling Less of More, Hyperion, 2008.
[10] W.B. Arthur, Competing technologies, increasing returns, and lock-in by historical events, Econ.J. 99 (394) (1989) 116–131.
[11] N. Abedini, S. Shakkottai, Content caching and scheduling in wireless networks with elastic and inelastic traffic, IEEE/ACM Trans. Netw. 22 (3) (2014) 864–874.
[12] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, R. Buyya, Cloud-based augmentation for mobile devices: motivation, taxonomies, and open challenges, IEEE Commun. Surv. Tutorials 16 (1) (2014) 337–368, https://doi.org/10.1109/SURV.2013.070813.00285.
[13] G.A. Akpakwu, B.J. Silva, G.P. Hancke, A.M. Abu-Mahfouz, A survey on 5g networks for the internet of things: communication technologies and challenges, IEEE Access 6 (2018) 3619–3647.
[14] A. Beloglazov, J. Abawajy, R. Buyya, Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing, Future Gener.Comput.Syst. 28 (5) (2012) 755–768.
[15] A.V. Banerjee, A simple model of herd behavior, Q. J. Econ. 107 (3) (1992) 797–817.
[16] F.M. Bass, A new product growth for model consumer durables, Manage. Sci. 15 (5) (1969) 215–227.
[17] A. Beloglazov, R. Buyya, Energy efficient resource management in virtualized cloud data centers, Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, IEEE Computer Society, 2010, pp. 826–831.
[18] A. Beloglazov, R. Buyya, Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers, Concurrency Comput 24 (13) (2012) 1397–1420.
[19] A.A. Bhattacharya, D. Culler, E. Friedman, A. Ghodsi, S. Shenker, I. Stoica, Hierarchical scheduling for diverse datacenter workloads, Proceedings of the 4th Annual Symposium on Cloud Computing, ACM, 2013, p. 4.
[20] L.A. Barroso, J. Clidaras, U. Hölzle, The datacenter as a computer: an introduction to the design of warehouse-scale machines, Synth.Lect.Comput.Archit. 8 (3) (2013) 1–154.
[21] D. Boyd, N.B. Ellison, Social network sites: definition, history, and scholarship, J. Comput.-Mediated Commun. 13 (1) (2007) 210–230, https://doi.org/10.1111/j.1083-6101.2007.00393.x.
[22] E. Berger, Dynamic monopolies of constant size, J. Comb. Theory Ser. B 83 (2) (2001) 191–200.
[23] D. Borthakur, J. Gray, J.S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, et al., Apache hadoop goes realtime at facebook, Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, ACM, 2011, pp. 1071–1080.
[24] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on Twitter, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, ACM, 2011, pp. 65–74.
[25] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an influencer: quantifying influence on Twitter, Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9–12, 2011, (2011), pp. 65–74, https://doi.org/10.1145/1935826.1935845.
[26] S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades, J.Political Economy (1992) 992–1026.
[27] D. Bottazzi, R. Montanari, A. Toninelli, Context-aware middleware for anytime, anywhere social networks, IEEE Intell. Syst. 22 (5) (2007).
[28] R. Bashroush, M. Noureddine, A cost effective cloud data centre capacity planning method based on modality cost analysis, Int. J. Commun. Netw.Distrib. Syst. 11 (3) (2013) 250–261.
[29] P. Bonacich, Power and centrality: a family of measures, Am.J.Sociol. (1987) 1170–1182.
[30] S.P. Borgatti, Centrality and network flow, Social Netw. 27 (1) (2005) 55–71, https://doi.org/10.1016/j.socnet.2004.11.008.
[31] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput.Netw. ISDN Syst. 30 (1) (1998) 107–117.
[32] E. Bakshy, I. Rosenn, C. Marlow, L.A. Adamic, The role of social networks in information diffusion, Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16–20, 2012, (2012), pp. 519–528, https://doi.org/10.1145/2187836.2187907.
[33] A. Brodersen, S. Scellato, M. Wattenhofer, YouTube around the world: geographic popularity of videos, Proceedings of the 21st international conference on World Wide Web, ACM, 2012, pp. 241–250.
[34] R. Bashroush, E. Woods, Architectural principles for energy-aware internet-scale applications, IEEE Softw. 34 (3) (2017) 14–17.
[35] Y. Chen, S. Alspaugh, R. Katz, Interactive analytical processing in big data systems: a cross-industry study of mapreduce workloads, Proc. VLDB Endowment 5 (12) (2012) 1802–1813.
[36] K. Chard, S. Caton, O. Rana, K. Bubendorfer, Social cloud: cloud computing in social networks, IEEE International Conference on Cloud Computing, CLOUD 2010, Miami, FL, USA, 5–10 July, 2010, (2010), pp. 99–106, https://doi.org/10.1109/CLOUD.2010.28.
[37] O. Coutand, O. Droegehorn, K. David, P. Nurmi, P. Floréen, R. Kernchen, S. Holtmanns, S. Campadello, T. Kanter, M. Martin, et al., Context-aware group management in mobile environments, IST Mobile Summit, (2005).
[38] S. Chib, E. Greenberg, Understanding the metropolis-hastings algorithm, Am. Stat. 49 (4) (1995) 327–335.
[39] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White paper, 2017.
[40] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, S.B. Moon, I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system, Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement 2007, San Diego, California, USA, October 24–26, 2007, (2007), pp. 1–14, https://doi.org/10.1145/1298306.1298309.
[41] Under the Hood: Scheduling MapReduce jobs more efficiently with Corona, Facebook engineering, 2012, (http://goo.gl/XJRNN). [Online; accessed 20-Dec-2017].
[42] P. Clifford, A. Sudbury, A model for spatial conflict, Biometrika 60 (3) (1973) 581–588.
[43] A.K. Dey, G.D. Abowd, D. Salber, A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications, Hum.-Comput.Interact. 16 (2) (2001) 97–166.
[44] C. Doerr, N. Blenn, S. Tang, P. Van Mieghem, Are friends overrated? A study for the social news aggregator digg. com, Comput. Commun. 35 (7) (2012) 796–809.

[45] Y. Dodge, D. Cox, D. Commenges, A. Davison, P. Solomon, S. Wilson, The Oxford Dictionary of Statistical Terms, Oxford University Press, USA, 2006.
[46] P. Delgado, D. Didona, F. Dinu, W. Zwaenepoel, Job-aware scheduling in eagle: divide and stick to your probes, Proceedings of the Seventh ACM Symposium on Cloud Computing, (2016). EPFL-CONF-221125
[47] P. Delgado, F. Dinu, A.-M. Kermarrec, W. Zwaenepoel, Hawk: hybrid datacenter scheduling, USENIX Annual Technical Conference, (2015), pp. 499–510.
[48] A.K. Dey, Understanding and using context, Pers Ubiquitous Comput. 5 (1) (2001) 4–7.
[49] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.
[50] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113.
[51] D. Daley, D.G. Kendall, Stochastic rumours, IMA J. Appl. Math. 1 (1) (1965) 42–55.
[52] C. Delimitrou, C. Kozyrakis, Paragon: Qos-aware scheduling for heterogeneous datacenters, ACM SIGPLAN Notices, 48 ACM, 2013, pp. 77–88.
[53] C. Delimitrou, C. Kozyrakis, Quasar: resource-efficient and qos-aware cluster management, ACM SIGPLAN Notices, 49 ACM, 2014, pp. 127–144.
[54] F.R. Dogar, T. Karagiannis, H. Ballani, A. Rowstron, Decentralized task-aware scheduling for data center networks, ACM SIGCOMM Computer Communication Review, 44 ACM, 2014, pp. 431–442.
[55] L. Dickens, I. Molloy, J. Lobo, P.-C. Cheng, A. Russo, Learning stochastic models of information flow, Data Engineering (ICDE), 2012 IEEE 28th International Conference on, IEEE, 2012, pp. 570–581.
[56] T.V.T. Duy, Y. Sato, Y. Inoguchi, Performance evaluation of a green scheduling algorithm for energy savings in cloud computing, Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on, IEEE, 2010, pp. 1–8.
[57] C. Delimitrou, D. Sanchez, C. Kozyrakis, Tarcil: reconciling scheduling speed and quality in large shared clusters, Proceedings of the Sixth ACM Symposium on Cloud Computing, ACM, 2015, pp. 97–110.
[58] D.A. Easley, J.M. Kleinberg, Networks, Crowds, and Markets - Reasoning About a Highly Connected World, Cambridge University Press, 2010.
[59] Ericsson research blog. 5g media delivery, 2017, (https://www.ericsson.com/research-blog/5g-media-delivery/). [Online; accessed 20-Dec-2017].
[60] E. Even-Dar, A. Shapira, A note on maximizing the spread of influence in social networks, Inf. Process. Lett. 111 (4) (2011) 184–187, https://doi.org/10.1016/j.ipl.2010.11.015.
[61] The Internet of Things: How the Next Evolution of the Internet is Changing Everything, CISCO, San Jose, CA, USA, 2011. White Paper
[62] A. Fard, A. Abdolrashidi, L. Ramaswamy, J.A. Miller, Towards efficient query processing on massive time-evolving graphs, 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2012, Pittsburgh, PA, USA, October 14–17, 2012, (2012), pp. 567–574, https://doi.org/10.4108/icst.collaboratecom.2012.250532.
[63] Facebook Newsroom, 2018, (http://newsroom.fb.com/Key-Facts). [Online; accessed 20-Dec-2017].
[64] J. Fowler, N. Christakis, Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives, HarperCollins Publishers, 2009.
[65] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, A. Akella, Multi-resource packing for cluster schedulers, ACM SIGCOMM Comput. Commun. Rev. 44 (4) (2015) 455–466.
[66] A. Goyal, F. Bonchi, L.V. Lakshmanan, Learning influence probabilities in social networks, Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 241–250.
[67] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, Proceedings of the 13th International Conference on World Wide Web, ACM, 2004, pp. 491–501.
[68] S. Govindan, J. Liu, A. Kansal, A. Sivasubramaniam, Cuanta: quantifying effects of shared on-chip resource interference for consolidated virtual machines, Proceedings of the 2nd ACM Symposium on Cloud Computing, ACM, 2011, p. 22.
[69] J. Goldenberg, B. Libai, E. Muller, Talk of the network: a complex systems look at the underlying process of word-of-mouth, Marketing Lett. 12 (3) (2001) 211–223.
[70] T. Gea, J. Paradells, M. Lamarca, D. Roldan, Smart cities as an application of Internet of Things: experiences and lessons learnt in barcelona, Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2013 Seventh International Conference On, IEEE, 2013, pp. 552–557.
[71] M. Granovetter, Threshold models of collective behavior, Am.j. Sociol. (1978) 1420–1443.
[72] I. Gog, M. Schwarzkopf, A. Gleave, R.N. Watson, S. Hand, Firmament: Fast, Centralized Cluster Scheduling at Scale, Usenix, 2016.
[73] A. García-Silva, J.-H. Kang, K. Lerman, O. Corcho, Characterising emergent semantics in Twitter lists, Proceedings of the 9th International Conference on The Semantic Web: research and applications (ESWC), Heraklion, Greece, (2012).
[74] A. Hinze, G. Buchanan, Context-awareness in mobile tourist information systems: challenges for user interaction, International Workshop on Context in Mobile HCI at the Seventh International Conference on Human Computer Interaction with Mobile Devices and Services, (2005).
[75] I. Hoque, I. Gupta, Disk layout techniques for online social network data, Internet Comput. IEEE 16 (3) (2012) 24–36.
[76] K. Henricksen, J. Indulska, A. Rakotonirainy, Infrastructure for pervasive computing: challenges, GI Jahrestagung (1), (2001), pp. 214–222.
[77] K. Huguenin, A.-M. Kermarrec, K. Kloudas, F. Taïani, Content and geographical locality in user-generated content sharing systems, Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video, ACM, 2012, pp. 77–82.
[78] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A.D. Joseph, R.H. Katz, S. Shenker, I. Stoica, Mesos: a platform for fine-grained resource sharing in the data center, NSDI, 11 (2011). 22–22
[79] R.A. Holley, T.M. Liggett, Ergodic theorems for weakly interacting infinite systems and the voter model, Ann.Probab. (1975) 643–663.
[80] T. Hoßfeld, R. Schatz, E. Biersack, L. Plissonneau, Internet video delivery in YouTube: from traffic measurements to quality of experience, Data Traffic Monitoring and Analysis, Springer, 2013, pp. 264–301.
[81] R. Istepanian, S. Hu, N. Philip, A. Sungoor, The potential of Internet of m-health Things "m-IoT" for non-invasive glucose level sensing, Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 5264–5266.
[82] Internet Society. Global Internet Report 2017. Mobile Evolution and Development of the Internet, 2017, (https://future.internetsociety.org/wp-content/uploads/2017/09/2017-Internet-Society-Global-Internet-Report-Paths-to-Our-Digital-Future.pdf). [Online; accessed 20-Dec-2017].
[83] M. Isard, V. Prabhakaran, J. Currey, U. Wieder, K. Talwar, A. Goldberg, Quincy: fair scheduling for distributed computing clusters, Proceedings of the ACM SIGOPS 22nd Symposium on Operating Systems Principles, ACM, 2009, pp. 261–276.
[84] E. Ising, Beitrag zur Theorie des Ferromagnetismus, Zeitschrift für Physik A Hadrons and Nuclei 31 (1) (1925) 253–258.
[85] International Telecommunication Union. ict facts and figures in 2017, 2017, (https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf). [Online; accessed 20-Dec-2017].
[86] F. Juarez, J. Ejarque, R.M. Badia, Dynamic energy-aware scheduling for parallel task-based application in cloud computing, Future Gener. Comput. Syst. (2016).
[87] L. Jiang, G. Feng, S. Qin, Content distribution for 5g systems based on distributed cloud service network architecture (2015).
[88] A. Jakóbik, D. Grzonka, J. Kołodziej, Security supportive energy aware scheduling and scaling for cloud environments (2017).
[89] V. Jacobson, D.K. Smetters, J.D. Thornton, M.F. Plass, N. Briggs, R. Braynard, Networking named content, Commun. ACM 55 (1) (2012) 117–124, https://doi.org/10.1145/2063176.2063204.
[90] M. Kuperman, G. Abramson, Small world effect in an epidemiological model, Phys. Rev. Lett. 86 (13) (2001) 2909–2912.
[91] R.T. Kaushik, M. Bhandarkar, Greenhdfs: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster, Proceedings of the USENIX Annual Technical Conference, (2010), p. 109.
[92] I. Kilanioti, Improving multimedia content delivery via augmentation with social information. the social prefetcher approach, Multimedia IEEE Trans. 17 (9) (2015) 1460–1470, https://doi.org/10.1109/TMM.2015.2459658.
[93] D. Kempe, J.M. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, ACM, 2003, pp. 137–146.
[94] E. Katz, P.F. Lazarsfeld, Personal Influence, The Part Played by People in the Flow of Mass Communications, Transaction Publishers, 1966.
[95] J.M. Kleinberg, Cascading behavior in networks: algorithmic and economic issues, in: N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani (Eds.), Algorithmic

Game Theory, Cambridge University Press, 2007, pp. 613–632.
[96] I. Kilanioti, G.A. Papadopoulos, Content delivery simulations supported by social network-awareness, Simul. Modell. Pract. Theory 71 (2017) 114–133.
[97] K. Karanasos, S. Rao, C. Curino, C. Douglas, K. Chaliparambil, G.M. Fumarola, S. Heddaya, R. Ramakrishnan, S. Sakalanaga, Mercury: hybrid centralized and distributed scheduling in large shared clusters, USENIX Annual Technical Conference, (2015), pp. 485–497.
[98] K. Lerman, S. Intagorn, J.-H. Kang, R. Ghosh, Using proximity to predict activity in social networks, Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 555–556.
[99] F. Liberal, A. Kourtis, J.O. Fajardo, H. Koumaras, Multimedia content delivery in sdn and nfv-based towards 5g networks, IEEE COMSOC MMTC E-Lett. 10 (4) (2015) 6–10.
[100] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst, Patterns of cascading behavior in large blog graphs, Proceedings of SIAM International Conference on Data Mining (SDM) 2007, SIAM, 2007.
[101] C.X. Lin, Q. Mei, Y. Jiang, J. Han, S. Qi, Inferring the diffusion and evolution of topics in social communities, Mind 3 (d4) (2011) d5.
[102] N.D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, A.T. Campbell, A survey of mobile phone sensing, IEEE Commun. Mag. 48 (9) (2010) 140–150, https://doi.org/10.1109/MCOM.2010.5560598.
[103] D. Liben-Nowell, J. Kleinberg, Tracing information flow on a global scale using internet chain-letter data, Proc. Natl. Acad. Sci. 105 (12) (2008) 4633–4638.
[104] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, S. Chouvardas, Placing dynamic content in caches with small population, IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications, (2016), pp. 1–9, https://doi.org/10.1109/INFOCOM.2016.7524380.
[105] T. Łuczak, Size and connectivity of the K-core of a random graph, Discrete Math. 91 (1) (1991) 61–68.
[106] X. Luo, Y. Wang, Z. Zhang, H. Wang, Superset: a non-uniform replica placement strategy towards high-performance and cost-effective distributed storage service, Advanced Cloud and Big Data (CBD), 2013 International Conference on, IEEE, 2013, pp. 139–146.
[107] Y.C. Lee, A.Y. Zomaya, Energy efficient utilization of resources in cloud computing systems, J. Supercomput. 60 (2) (2012) 268–280.
[108] A. Menon, Big data@ facebook, Proceedings of the 2012 Workshop on Management of Big Data Systems, ACM, 2012, pp. 31–32.
[109] G. Maier, A. Feldmann, V. Paxson, M. Allman, On dominant characteristics of residential broadband internet traffic, Proceedings of the 9th ACM SIGCOMM Internet Measurement Conference IMC, ACM, 2009, pp. 90–102.
[110] S. Milgram, The Small World problem, Psychol.Today 2 (1) (1967) 60–67.
[111] Mobile-Edge Computing – Introductory technical white paper, 2018, https://goo.gl/ybrCnq Accessed: 2016-03-15.
[112] S. Morris, Contagion, Rev. Econ. Stud. 67 (1) (2000) 57–78.
[113] J. Mars, L. Tang, Whare-map: heterogeneity in homogeneous warehouse-scale computers, ACM SIGARCH Computer Architecture News, 41 ACM, 2013, pp. 619–630.
[114] N.D.A.B. Navarro, C.A. Da Costa, J.L.V. Barbosa, R.D.R. Righi, A context-aware spontaneous mobile social network, Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015 IEEE 12th Intl Conf on, IEEE, 2015, pp. 85–92.
[115] A. Najar, L. Denoyer, P. Gallinari, Predicting information diffusion on social networks with partial knowledge, Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 1197–1204.
[116] V. Nejković, F. Jelenković, M. Tošić, N. Milošević, Z. Nikolić, Coordss: an ontology framework for heterogeneous networks experimentation, Telfor J. 8 (2) (2016) 70–75.
[117] R. Nathuji, A. Kansal, A. Ghaffarkhah, Q-clouds: managing performance interference effects for qos-aware clouds, Proceedings of the 5th European conference on Computer systems, ACM, 2010, pp. 237–250.
[118] M. Nekovee, Y. Moreno, G. Bianconi, M. Marsili, Theory of rumour spreading in complex social networks, Physica A 374 (1) (2007) 457–470.
[119] K. Ousterhout, P. Wendell, M. Zaharia, I. Stoica, Sparrow: distributed, low latency scheduling, Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, ACM, 2013, pp. 69–84.
[120] V. Pejovic, M. Musolesi, Anticipatory mobile computing: a survey of the state of the art and research challenges, ACM Comput. Surv. 47 (3) (2015) 47, https://doi.org/10.1145/2693843.
[121] L. Plissonneau, G. Vu-Brugier, Mobile data traffic analysis: how do you prefer watching videos? Proceedings of the 22nd International Teletraffic Congress (ITC), IEEE, 2010, pp. 1–8.
[122] C. Perera, A. Zaslavsky, P. Christen, D. Georgakopoulos, Context aware computing for the internet of things: a survey, IEEE Commun. Surv. Tutorials 16 (1) (2014) 414–454.
[123] C. Perera, A. Zaslavsky, P. Christen, D. Georgakopoulos, Sensing as a service model for smart cities supported by internet of things, Trans. Emerging Telecommun.Technol. 25 (1) (2014) 81–93.
[124] H. Qu, O. Mashayekhi, D. Terei, P. Levis, Canary: a scheduling architecture for high performance cloud computing, arXiv:1602.01412v1(2016).
[125] S. Ricciardi, D. Careglio, J. Sole-Pareta, U. Fiore, F. Palmieri, et al., Saving energy in data center infrastructures, Data Compression, Communications and Processing (CCP), 2011 First International Conference on, IEEE, 2011, pp. 265–270.
[126] K. Ren, Y. Kwon, M. Balazinska, B. Howe, Hadoop's adolescence: an analysis of hadoop usage in scientific workloads, Proc. VLDB Endowment 6 (10) (2013) 853–864.
[127] J. Rasley, K. Karanasos, S. Kandula, R. Fonseca, M. Vojnovic, S. Rao, Efficient queue management for cluster scheduling, Proceedings of the Eleventh European Conference on Computer Systems, ACM, 2016, p. 36.
[128] C. Ren, E. Lo, B. Kao, X. Zhu, R. Cheng, On querying historial evolving graph sequences, Proc. VLDB Endowment 4 (11) (2011).
[129] M.G. Rodriguez, J. Leskovec, B. Schölkopf, Structure and dynamics of information pathways in online media, Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy, (2013).
[130] E.M. Rogers, Diffusion of innovations, Simon and Schuster, 1995.
[131] C. Reiss, A. Tumanov, G.R. Ganger, R.H. Katz, M.A. Kozuch, Heterogeneity and dynamicity of clouds at scale: Google trace analysis, Proceedings of the Third ACM Symposium on Cloud Computing, ACM, 2012, p. 7.
[132] C. Reiss, J. Wilkes, J.L. Hellerstein, Google cluster-usage traces: format + schema, Technical Report, Google Inc., Mountain View, CA, USA, 2011. Revised 2012.03.20. Posted at http://code.google.com/p/googleclusterdata/wiki/TraceVersion2
[133] M. Satyanarayanan, Pervasive computing: vision and challenges, IEEE Pers. Commun. 8 (4) (2001) 10–17.
[134] X. Song, Y. Chi, K. Hino, B.L. Tseng, Information flow modeling based on diffusion rate for prediction and ranking, Proceedings of the 16th International Conference on World Wide Web, ACM, 2007, pp. 191–200.
[135] T.C. Sendling, Micromotives and Macrobehavior, New York and London: Norton, 1978.
[136] D. Shue, M.J. Freedman, A. Shaikh, Performance isolation and fairness for multi-tenant cloud storage, OSDI, 12 (2012), pp. 349–362.
[137] H. Sundmaeker, P. Guillemin, P. Friess, S. Woelfflé, Vision and challenges for realising the internet of things, Cluster Eur. Res. Projects Internet Things Eur. Commision 3 (3) (2010) 34–36.
[138] M. Schwarzkopf, A. Konwinski, M. Abd-El-Malek, J. Wilkes, Omega: flexible, scalable schedulers for large compute clusters, Proceedings of the 8th ACM European Conference on Computer Systems, ACM, 2013, pp. 351–364.
[139] B.N. Schilit, A. LaMarca, G. Borriello, W.G. Griswold, D. McDonald, E. Lazowska, A. Balachandran, J. Hong, V. Iverson, Challenge: ubiquitous location-aware computing and the place lab initiative, Proceedings of the 1st ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots, ACM, 2003, pp. 29–35.
[140] The Smart Grid: An Introduction. US Department of Energy, 2008, http://goo.gl/jTNgf Accessed: 2016-03-15.
[141] S. Scellato, C. Mascolo, M. Musolesi, J. Crowcroft, Track globally, deliver locally: improving Content Delivery Networks by tracking geographic social cascades, Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28, - April 1, 2011, (2011), pp. 457–466, https://doi.org/10.1145/1963405.1963471.

[142] B.N. Schilit, M.M. Theimer, Disseminating active map information to mobile hosts, IEEE Netw. 8 (5) (1994) 22–32.
[143] S. Sohrabi, A. Tang, I. Moser, A. Aleti, Adaptive virtual machine migration mechanism for energy efficiency, Proceedings of the 5th International Workshop on Green and Sustainable Software, ACM, 2016, pp. 8–14.
[144] N. Sastry, E. Yoneki, J. Crowcroft, Buzztraq: predicting geographical access patterns of social cascades using social networks, Proceedings of the Second ACM EuroSys Workshop on Social Network Systems, SNS 2009, Nuremberg, Germany, March 31, 2009, (2009), pp. 39–45, https://doi.org/10.1145/1578002.1578009.
[145] E. Thereska, A. Donnelly, D. Narayanan, Sierra: practical power-proportionality for data center storage, Proceedings of the sixth conference on Computer systems, ACM, 2011, pp. 169–182.
[146] R. Torres, A. Finamore, J.R. Kim, M. Mellia, M.M. Munafo, S. Rao, Dissecting video server selection strategies in the youtube cdn, Distributed Computing Systems (ICDCS), 2011 31st International Conference on, IEEE, 2011, pp. 248–257.
[147] S. Traverso, K. Huguenin, I. Trestian, V. Erramilli, N. Laoutaris, K. Papagiannaki, TailGate: handling long-tail content with a little help from friends, Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16–20, 2012, (2012), pp. 151–160, https://doi.org/10.1145/2187836.2187858.
[148] M. Tosic, V. Nejkovic, F. Jelenkovic, N. Milosevic, Z. Nikolic, N. Makris, T. Korakis, Semantic coordination protocol for lte and wi-fi coexistence, Networks and Communications (EuCNC), 2016 European Conference on, IEEE, 2016, pp. 69–73.
[149] Management of Networks with Constrained Devices: Use Cases. IETF Internet Draft, 2015, https://goo.gl/cT5pXr Accessed: 2016-03-15.
[150] C. Van den Bulte, Y.V. Joshi, New product diffusion with influentials and imitators, Marketing Sci. 26 (3) (2007) 400–421.
[151] V.K. Vavilapalli, A.C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, et al., Apache hadoop yarn: Yet another resource negotiator, Proceedings of the 4th annual Symposium on Cloud Computing, ACM, 2013, p. 5.
[152] E.J. Vergara, S. Nadjm-Tehrani, Energybox: a trace-driven tool for data transmission energy consumption studies, Energy Efficiency in Large Scale Distributed Systems - COST IC0804 European Conference, EE-LSDS 2013, Vienna, Austria, April 22–24, 2013, Revised Selected Papers, (2013), pp. 19–34, https://doi.org/10.1007/978-3-642-40517-4_2.
[153] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, J. Wilkes, Large-scale cluster management at google with borg, Proceedings of the Tenth European Conference on Computer Systems, ACM, 2015, p. 18.
[154] D.J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. 99 (9) (2002) 5766–5771.
[155] R. Want, T. Pering, System challenges for ubiquitous & pervasive computing, Proceedings of the 27th international conference on Software engineering, ACM, 2005, pp. 9–14.
[156] Z. Wang, X. Zhou, Z. Yu, H. Wang, H. Ni, Quantitative evaluation of group user experience in smart spaces, Cybern. Syst. 41 (2) (2010) 105–122.
[157] H. Yang, A. Breslow, J. Mars, L. Tang, Bubble-flux: precise online qos management for increased utilization in warehouse scale computers, ACM SIGARCH Computer Architecture News, 41 ACM, 2013, pp. 607–618.
[158] J. Yang, J. Leskovec, Modeling information diffusion in implicit networks, Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), IEEE, 2010, pp. 599–608.
[159] YouTube statistics, 2018, (https://www.youtube.com/yt/press/statistics.html). [Online; accessed 20-Dec-2017].
[160] Y. Yan, Y. Qian, H. Sharif, D. Tipper, A survey on smart grid communication infrastructures: motivations, requirements and challenges, Commun. Surv. Tutorials IEEE 15 (1) (2013) 5–20.
[161] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, I. Stoica, Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling, Proceedings of the 5th European Conference on Computer systems, ACM, 2010, pp. 265–278.
[162] X. Zhang, E. Tune, R. Hagmann, R. Jnagal, V. Gokhale, J. Wilkes, Cpi 2: Cpu performance isolation for shared compute clusters, Proceedings of the 8th ACM European Conference on Computer Systems, ACM, 2013, pp. 379–391.
[163] D. Zhang, Z. Yu, B. Guo, Z. Wang, Exploiting personal and community context in mobile social networks, Mobile Social Networking, Springer, 2014, pp. 109–138.
[164] F. Zhou, L. Zhang, E. Franco, A. Mislove, R. Revis, R. Sundaram, WebCloud: recruiting social network users to assist in content distribution, Proceedings of the 11th IEEE International Symposium on Network Computing and Applications, Cambridge, MA, USA, (2012).