IT and the Consumer

3

In this chapter we look at the implications for consumers in light of the developments (technological and otherwise) that we have presented in the previous two chapters. We start with electronic commerce and look at various approaches that businesses, both brick-and-mortar and mobile, now have at their disposal for getting their message to the customer, including advertising, social media marketing, and recommendation. We will also show how big data analytics can open up new possibilities. Finally, the emerging area of e-government will be touched upon. As will be seen, we here take a customer's perspective and argue that not all of the above is beneficial, since there are pros and cons of the various ways customers are nowadays addressed or approached, and there are also many things that they not only now *can*, but indeed *have* to do themselves.

3.1 Commercialization of the Web

Electronic commerce or *e-commerce* for short is generally understood as the process of buying, selling, transferring, or exchanging products, services, or information via computer networks; in particular via the Web. It has become remarkably popular over the past 25 years since it provides an approach for conducting business that was highly innovative in the beginning due to its potential global reach, something that is difficult for traditional brick-and-mortar businesses to imitate. Furthermore it represents a reduction in the cost of transactions, it can provide unique, customized products for even small customer bases, and it allows customer access 24 hours a day, 7 days a week, 365 days a year ("24/7/365"). Moreover, e-commerce *between* customers, or C2C, has been highly popularized through auctioning sites like eBay or TradeMe. Often, a fundamental characteristic of an e-commerce scenario is the absence of *intermediaries*, i.e., third parties offering intermediation services to two trading parties. For example, a publishing company can now sell directly to readers, without going through the intermediary of a book

G. Vossen et al., *The Web at Graduation and Beyond*,
DOI 10.1007/978-3-319-60161-8_3

store, authors can directly sell to readers without a publisher as intermediary, travelers can circumvent travel agencies, or individuals can sell used cars without the involvement of a car dealer.

3.1.1 Components of an E-Commerce System

A typical e-commerce system has the following major components:

- **Product presentation component**: An e-commerce system must provide various ways for customers (including businesses) to search, select, compare, and identify products they want to purchase. This presentation component typically needs to service several channels, including browsers of various makes and sizes running on a variety of distinct devices or platforms; the presentation itself nowadays needs what is called *responsive design* so that is can adapt itself to the various screen sizes in use today.
- **Inventory management and supply chain component**: When a customer orders, the e-commerce system needs to provide an immediate response on whether the desired product, item, or service is currently available and, if not, how long it might take to restock or rebuild. Conversely, the system should be capable of efficient warehouse management and be able to automate the interaction of a business with its suppliers to a great extent.
- **Order entry and shopping basket component**: After the customer has made a selection, he or she needs to enter an order for the product into the electronic commerce system. Order entry often allows the customer to add items to an electronic shopping basket, which is a list of the products the customer has selected to purchase. Before an item is added to that basket, the e-commerce system should have the inventory control system check the product database to see if there is adequate stock on hand or if the product needs to be ordered from a manufacturer.
- **Payment component**: To allow customers to pay for the items purchased, an e-commerce system needs to have electronic payment capabilities. Various approaches are used for electronic payment, including payment by credit card, third-party payment service (e.g., PayPal) or by electronic funds transfer. To ensure the security of the card information sent over the Internet, special protocols such as HTTPS that provide data encryption are used. Interestingly, it has become common to ask for prepayment from a customer (during checkout), a habit that was less common prior to electronic commerce (and still is in many classical business fields).
- **Customer service and support component**: At any time before, during, or after purchasing an item, the customer may need advice or additional services or support. For example, a customer may have a question about how a particular item of clothing fits before purchasing it, or the customer may have a question regarding delivery. After receiving an item, the customer may decide to exchange or return the item. Sometime later, the customer may have a warranty

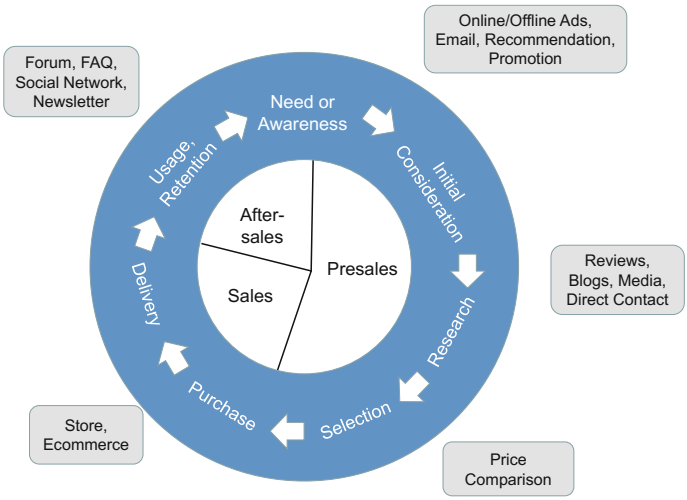

Fig. 3.1 Phases of a customer journey

claim. Many of these customer service situations can be dealt with by providing detailed information and answers to questions (in FAQs) electronically, and by providing appropriate customer services. The general perception and goal today is that of providing a smooth *customer journey* (or *customer experience*, CX, see Chap. 4) that starts with producing awareness for a product or service and continues through the phases of consideration, purchase, retention, and even advocacy as shown in Fig. 3.1. This journey typically alternates between a variety of physical and digital touchpoints and is nowadays supported by comprehensive analytics of all the data traces a customer leaves behind.

- **Recommendation component**: Closely related to the previous point (and in particular the awareness and consideration phases) is that of recommendation, which has become popular in e-commerce, as people can now refer to other people for obtaining advice on a product or service. As will be seen ion Sect. 3.6, an e-commerce site is typically interested in establishing a "profile" for every customer containing, for example, all the products the customer has ever bought or at least looked at. It gets interesting when the site is able to find other customers whose profile is, in a sense that needs to be made precise, "similar", so the products the first customer has not yet acquired can be recommended to her or him. Known as collaborative filtering, this scheme applies to goods, movies, music, and even people on a dating site.

Figure 3.2 summarizes the various components an e-commerce system must have today, where core components are shown in blue and interfaces to internal as

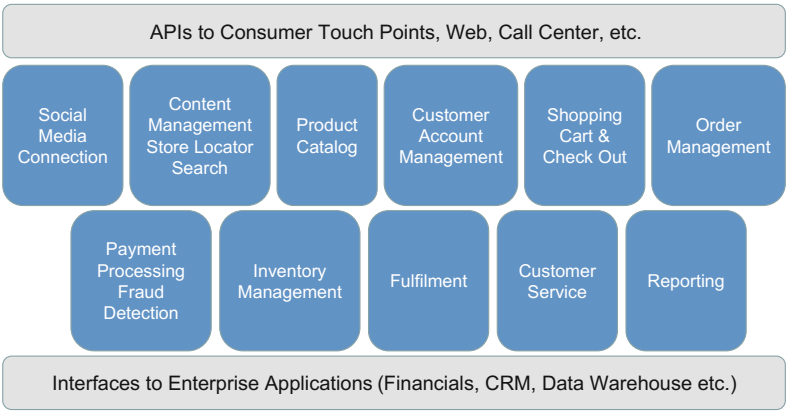

Fig. 3.2 E-commerce components

well as external applications in green. The customer (although not explicitly shown) and customer service play a central role, as do social media.

Clearly, while these components are well-understood these days, it has taken more than 20 years of development and (gathering of) experience to develop this understanding. In the beginning, i.e., in the mid- to late-1990s, user acceptance of e-commerce was low, due to limitations in Internet access, the limited number of companies doing e-business at all, a general lack of trust, or due to the missing customer support. For further details on the issues of the early days and the development until today, the interested reader should consult, for example, consecutive reports by Mary Meeker of Silicon Valley venture capital company KPCB.[1] Indeed, the initial perception was that a traditional retail shop will know returning customers after a short while, whereas an e-commerce site, without further measures, cannot distinguish an HTTP request by a customer today from one by the same customer tomorrow. This situation has meanwhile changed considerably, and it is one of many situations where big data has become prominent: E-businesses today typically know their customers well, can classify them according to a variety of criteria, and are often even able to predict what the next transaction or the next step during the respective customer journey will be.

It took e-commerce companies some time to recognize that doing their business electronically involves much more than setting up a Web site that comprises the components listed above. Indeed, it requires a considerable amount of process reengineering, as, for example, the Web shop front and the back office must be connected in ways that did not exist before; this was frequently overlooked in the early days of electronic commerce and was a typical cause of failure. Among the fastest to realize this and to react appropriately have been banks and financial

[1] see www.kpcb.com/blog/2016-internet-trends-report for the latest edition.

institutions, since their business has moved to the Web quite extensively; electronic banking, stock trade as well as online fund and portfolio management are in wide use today. A general theme (and challenge) in this context is the broad *digitization* of business, a topic that we will discuss in Chap. 5.

The move from stationary commerce to electronic and later to mobile commerce has also triggered the development of new branches of the software industry, which not only provide shop front software, but also systems for click stream analysis, payment encryption (including technologies such as Blockchain and parallel currencies such as Bitcoin), data mining, and customer relationship management (CRM). As we mentioned in Chap. 2, data mining has become popular for analyzing the vast amounts of data that are aggregated by a typical e-commerce installation; prominent data mining applications include association rule determination, clustering, and classification, some of which will be discussed in Chap. 4. Click streams that have been created by users are also subject to intensive data mining, and CRM comprises a set of tools and techniques for exploiting data mining results in order to attract new customers or to retain existing ones.

An e-commerce platform typically serves both the buyer as well as the seller side and sometimes even an intermediary. On the buyer side, there is a number of suppliers from which the company obtains its raw materials or supplies, often through some form of procurement process, and, thanks to the flattening we have mentioned in Chap. 1, this process can be executed world-wide. Internally, there is a supply-chain management (SCM) system at work that interacts with an enterprise resource planning (ERP) system in various ways. On the seller side, there may be several channels through which the company sells its goods or services also world-wide, including individual consumers, businesses, and partners (see also Fig. 3.1). For their various customers, some form of CRM-system will be in place in order to take care of after-sales activities, customer contacts, complaints, warranty claims, help desk inquiries, etc.

3.1.2 Types of Electronic Commerce

Electronic commerce has developed into various types that all have their specific properties and requirements:

- Business-to-consumer or B2C e-commerce, which involves a business selling its goods or services electronically to end-consumer.
- Business-to-business or B2B e-commerce, which involves a business selling its goods or services electronically to other businesses.
- Consumer-to-consumer or C2C e-commerce, which involves a consumer selling goods or a service electronically to other consumers.

B2C e-commerce is probably best known to the general public, although by figures alone B2B is considerably higher in value of goods traded. B2C e-commerce has become popular due to two aspects: the end of intermediaries and better price transparency. Indeed, goods are now often sold directly by a business to the end-consumer, instead of going through a third-party in between. For example, software can now be downloaded from the producer directly (called *electronic software distribution*), instead of having it delivered on DVD and sold through stores. As a result, prices may be lower (an expectation often not valid), or the seller will make a better profit (since nothing needs to be paid to the intermediary). Second and as mentioned, it has become very popular on the Web to provide price comparisons, through sites such as dealtime.com or guenstiger.de; as a consequence, the consumer will nowadays often do extensive comparisons before committing to a particular seller. We mention that B2C e-commerce has meanwhile reached almost every kind of goods; what started with books, music, and movies has nowadays reached even, for example, raw as well as processed or fresh food. One of the enablers for the latter has been the fact that delivery has been perfected over the years, and the nowadays same-day delivery is an option in many places around the globe. And while even a few hours might be too long a duration a customer has to wait for a product just ordered, Amazon is experimenting with delivery by drones ("Prime Air") that can bring delivery time—currently just in selected areas—down to a few minutes from the receipt of an order.[2]

B2B e-commerce, in turn, comes in three major varieties: In a *supplier-oriented marketplace*, a supplier provides e-commerce capabilities for other businesses to order its products; the other businesses place orders electronically from the supplier, much in the same way that consumers will place orders in B2C e-commerce. In a *buyer-oriented marketplace*, the business that wants to purchase a product requests quotations or bids from other companies electronically; each supplier that is interested places a bid electronically and the buyer selects the winning supplier from the submitted bids; this type is common in the car industry, where manufacturers requests bids from, say, tire suppliers. Finally, in an *intermediary-oriented marketplace*, a third-party business acts as an intermediary between the supplier and the buyer; the intermediary provides e-commerce capabilities for both suppliers and buyers in order to identify each other and to electronically transact business.

The third major type of e-commerce, C2C, is mostly manifested these days through auctions such as eBay or TradeMe.co.nz. eBay (Cohen 2002) was founded in September 1995 by computer programmer Pierre Omidyar under the name AuctionWeb. One of the early items sold on eBay was Omidyar's broken laser pointer, which to his surprise was due to a real interest in such an item by the winning bidder. The company officially changed the name of its service from AuctionWeb to eBay in September 1997. Millions of collectibles, appliances,

[2]see www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011 for a December 2016 experiment in this direction.

computers, furniture, CDs, DVDs, musical instruments, diecast models, outdoor equipment, cars, and other items are listed, bought, and sold daily on eBay. Some items are rare and valuable, while many other items would have been discarded if eBay with its thousands of bidders worldwide would not exist. Anything used or new can be sold as long as it is not illegal or does not violate the eBay *Prohibited and Restricted Items policy*. Interestingly, programmers can create applications that integrate with eBay through the eBay application programming interface (API) by joining the eBay Developers Program. This opens the door to "mashing up" eBay with totally different applications, an aspect we will have to say more about later.

We note that other types of e-commerce have meanwhile emerged, including G2C (government-to-citizen) or B2G (business-to-government), to name just two, and each type can be combined with one or more business models in order to actually generate revenue. All of these are characterized by two major features, which are *automation* and *self-service*.

Automation has been a major goal for the establishment of electronic commerce, as many of the interactions between the components that we saw in Fig. 3.2 are performed algorithmically. For example, recommendations of products to customers are determined based on the customer's buying history as well as on that of "related" or "similar" customers, where similarity can be based on various measures, all of which allow for automated computation.

Self-service, on the other hand, has always been promoted as a major convenience feature of electronic commerce. The customer is now "enabled" to book flights and hotels himself, to check whether a package has been shipped and is hence out for delivery, or to manage everything related to his or her bank account in a do-it-yourself fashion. While it has indeed become convenient to be able to access, for example, all information relating to air travel (including ticket prices, seat assignments, airline safety, or flight progress) by yourself, what has actually happened is that, as noted by Tom Friedman 10 years ago, the customer has in fact become an airline employee who works for free. This is even more apparent in banks, which are not only eliminating one branch after another, but also make an account holder pay if he or she needs manual help and wants a bank clerk to perform a money transfer! We will come back to this DIY aspect of e-commerce when we discuss disruption and other aspects in later chapters.

3.1.3 Recommendation, Advertising, Intermediaries

For anyone interested in setting up an e-commerce site or to embark on e-commerce in another way, two questions will prevail: What is an appropriate way to sell my product, and how do I attract traffic to my site? Especially for physical goods, it has become very helpful if such *advice comes from other customers*. On places like Amazon, eBay and others, this has become one of the main aspects people seek when shopping for a product: what others have said about that product, how much

they like it, whether or not they would buy it again, and maybe how the seller performed or how easy the overall experience has been. Once the importance of other customers' opinions, evaluations, and recommendations had been recognized, many Web shop providers started to install facilities for commenting on a regular and intensive basis. Amazon was among the first to take this even a step further and go from pure reviews, which can be commented by others, to a comprehensive recommendation system ("Customers who bought this item also bought …"), whose goal is, as an Amazon employee once put it, "to make people buy stuff they did not know they wanted." Integration of social media established such retailers as proponents of what is now known as *social* commerce.

The other aspect the owner of an e-commerce business will be interested in, the attraction of traffic as already mentioned above in the context of portals, is closely related to classical disciplines from business administration: advertising and marketing. Traditionally, *advertising* is the art of drawing public attention to goods or services by promoting a business, and is performed through a variety of media. It is a mechanism of *marketing*, which is concerned with the alignment of corporations with the needs of the business market (the term *customer journey* we mentioned above is also a term commonly used in marketing). We will take a closer look at online advertising below.

While we have mentioned that, from a seller's perspective, doing business over the Web may be attractive due to the absence of intermediaries which often do little more that take a share of the profit, there are situations in which new intermediaries enter the picture and indeed offer a highly valued service. This in particular refers to facilitating payments, for which *trusted third parties* have come onto the scene. The term is borrowed from cryptography, where it denotes an entity enabling interactions between two parties who both trust the third party; so that they can utilize this trust to secure their business interactions. A well-known example is PayPal, which allows payments and money transfers to be made over the Web, actually to anybody with an email address, and it performs payment processing for online vendors, auction sites, and other corporate users, for which it charges a fee. Private users need to register and set up a profile which, for example, includes a reference to a bank account or to a credit card; PayPal will use that reference to collect money to be paid by the account or card owner to someone else. When a customer reaches the checkout phase during an e-commerce session, the shopping site he or she interacts with might direct them to PayPal for payment processing. If the customer agrees to pay through PayPal, PayPal will verify the payment through a sequence of encrypted messages; if approved, the seller will receive a message stating that the payment has been verified, so that the goods can finally be shipped to the customer. Services such as PayPal or Escrow have invented the notion of a *micro-payment*, i.e., a tiny payment (of often just a few cents) which is feasible as a reimbursement only if occurring sufficiently often. Figure 3.3 shows the checkout process when the PayPal "Buy Now" button is employed.

Fig. 3.3 Checkout process utilizing PayPal Buy Now button

In conclusion, it is fair to say that since the inception of the Web in 1993, a considerable amount of the trading and retail business has moved to electronic platforms and is now run over the Web. E-commerce continues to grow at a fast pace; indeed, US e-commerce retail alone grew from $132 Billion in 2008 to a estimated $224 Billion in 2014. More precise figures can be found, for example, at www.emarketer.com/. This has created completely new industries, and it has led to a number of new business models and side effects that do not exist, at least not at this scale, in the physical world.

3.1.4 Case Amazon

Before we continue our elaboration of IT and its impact on the consumer, we pause for a moment and take a look at the "role model" for electronic commerce: Amazon.com. Amazon sold its first book online in July 1995 and, as mentioned in Chap. 1, soon after started selling CDs, then DVDs, and then items in many other categories. According to an article in USA Today,[3] "now you can hop online from your phone, download the e-book version, bid on a vintage couch on which to read it, and hire someone to explain the concepts to you—all with one click."

[3] www.usatoday.com/story/news/nation-now/2015/07/14/working—amazon-disruptions-timeline/30083935/

In the 20 years since its inception, Amazon has introduced major innovations that have had an impact on how people read, consume music and movies, and even shop for typical consumer products such as clothes and groceries. Beyond this, Amazon has invented (or at least experimented with) novel ways of delivering goods (e.g., by drones), and has become one of the largest cloud providers worldwide. As we have discussed in Chap. 2, Amazon is nowadays providing all of infrastructure, platform, and application software as service in the cloud. Not surprisingly, the Internet traffic caused by AWS has long (actually in January 2007) bypassed the traffic caused by Amazon's e-commerce business, as can be seen from media.amazonwebservices.com/blog/2008/big_aws_bandwidth.gif.

More importantly, Amazon has become the role model of a "disruptor" in the sense that, even without a single physical bookstore, Amazon has been able to disrupt the physical book market and its value network, and replace it by a virtual market that is cheaper, faster, and way more convenient than the classical market in many respects. We will discuss disruption and disruptive innovation in more detail in Chap. 5, but a brief summary of "Case Amazon" is presented here.

We follow the USA Today article mentioned above from July 14, 2015 on Amazon's lifestyle innovations, which summarizes the following key areas of innovation:

- Bookselling: In the mid-1990s, Amazon was the first to try online bookselling, and it succeeded widely and even put competitors (i.e., Borders) out of business.
- One-click purchasing: Buying something with one click only was introduced in the fall of 1997; Amazon even holds a patent for this.
- The cloud: We discussed AWS in Chap. 2, an idea that started in 2006 and has evolved into a major Amazon business.
- The cloud at home: In March 2015, the company expanded its professional services marketplace, Amazon Local Services. Now you can hire anything from a plumber to a goat herder, at your digital leisure.
- Members only: Amazon Prime was created in 2005, with users paying a flat annual subscription fee for certain benefits, including one-day shipping prices. Same-day delivery for Prime members launched in May 2015. Amazon announced its goals to launch drone delivery called Amazon Prime Air in December 2013, and said that the future delivery system is "designed to safely get packages into customers' hands in 30 minutes or less using small airborne devices".
- The rise of e-books and e-reading: Amazon's Kindle was a game-changer for e-reading. Launched in 2007, the Kindle connected book purchasing with book platforms, leading to sensational headlines like "print is dead." In mid-2010, Amazon's Kindle e-book sales outpaced hardcover book sales for the first time. By a similar token, the Kindle Fire, launched in September 2011, was a cheap version of the flashier Apple counterparts. Beyond this, Amazon purchased Goodreads, the leading social network for book lovers, in March 2013.

- Publishing: If you're selling, why not publish? In 2009, Amazon Publishing took the publishing world by storm, slowly snatching up titles from publishers such as Random House, Penguin, Macmillan, or Simon & Schuster.
- Audio: Amazon purchased audiobook mainstay Audible for $300 million in 2008, so there is a good chance it is powering all the audio books one will ever download.
- Robots: For the 2014 holiday rush, Amazon hired 15,000 robots for its fulfillment system.

Currently we even see the arrival of brick-and-mortar Amazon stores ("Amazon Books") in the US (see Sect. 5.5). Amazon has also been instrumental in the development of search, data mining, and recommendation technology, much of which is nowadays subsumed under "big data technology." This technology includes supply chain optimization (e.g., site selection for warehouses to minimize distribution costs; selection of optimal routes, schedules, and products groupings, to minimize delivery costs; minimization of time spent by drivers in traffic jams), pricing and profit optimization, fraud detection for credit card transactions or detection of criminal behavior on AWS (system intrusions, hacking attempts or other malicious activity), fake reviews detection, search engine technology to help users find what they want to buy quickly, customer segmentation, churn analysis, inventory and sales forecasting, payments analytics (for authors, vendors, publishers), competitive analysis, i.e., automatic processing and analysis of billions of comments posted by users on social networks about Amazon, its competitors, and new trends and taking action based on the respective findings.

Analyzing reviews is a particularly tricky area, since the majority of customers are either not willing to provide feedback or do not have the time to do so. As Suw Charman-Anderson wrote in Forbes in 2012,[4] "Amazon's reviews system is fundamentally broken and whilst that might seem like an issue that troubles only those of us in the industry who pay attention to these things, it isn't. As book reviews become more and more unreliable, so more and more buyers will start to get frustrated that they aren't getting what they were expected and will start looking for reviews elsewhere. That will habituate them to looking outside Amazon for information on books and bring Amazon's position as the canonical reference for books under threat." Indeed, fake reviews have become a lucrative business, and they often do not come from the manufacturer of a product as might be expected, but from a company that has been hired to produce such a review. There is already a name for this business, *astroturfing*, and legal instances are increasingly involved in fighting it, including New York Attorney General Eric T. Schneiderman, who made local businesses pay a total of $350,000 in fines for engaging in this illegal practice.[5]

[4] www.forbes.com/sites/suwcharmananderson/2012/12/18/amazon-is-ripe-for-disruption/#55ad91947d4c

[5] www.entrepreneur.com/article/228525

On the other hand, Amazon has been instrumental over the past 20 years in the development of data mining techniques that take advantage of the massive amount of digital data that is available on an e-commerce site, including customers' searches, clicks paths, length of stay on a particular site, buying histories, wish lists etc. Every single click is recorded and eventually analyzed, which has led to features such as "Customers Who Bought This Item Also Bought" or "Customers Who Viewed This Item Also Viewed." While content-based recommendations as well as collaborative filtering are commonly employed, there remains room for improvement in this space.

3.2 Big Data Analytics Application Areas

In Chap. 2 we discussed technology for the management of big data, and we gave two brief examples (matrix-vector multiplication and weather data analysis) for an application of the map-reduce paradigm. But what are some actual cases where big data has resulted in a new insight and how does this technology benefit business in general? In this section we will describe several application areas which demonstrate how big data can indeed be considered a game changer, since it has already led to developments that differ significantly from what we have seen in the past, e.g., in the context of data warehouses. These areas exhibit great variety, so it is important to note that our sample is not exhaustive and also represents only the beginning of the development, generally termed "big data analytics."

Sports

One of the oldest examples of what data can do, before it was called "big" data and be popularized, is from the area of sports and relates to the Oakland Athletics baseball team and their coach Billy Beane, who was able to use statistics and player data to revamp the team from an unsuccessful one into a comparatively successful one within a relatively short time span. The story is well documented in the book by Lewis (2004) and in a 2011 movie based on that book (with Brad Pitt as main actor). Doug Laney, whom we already mentioned in Sect. 2.3, provided a more recent example from sports on his blog, namely from the Indy 500 race happening in the U. S. every year on Memorial Day weekend. According to Laney, a present-day Indy 500 race car is on the inside "smattered with nearly 200 sensors constantly measuring the performance of the engine, clutch, gearbox, differential, fuel system, oil, steering, tires, drag reduction system (DRS), and dozens of other components, as well as the drivers' health. These sensors spew about 1 GB of telemetry per race to engineers pouring over them during the race and data scientists crunching them between races. According to McLaren, its computers run a thousand simulations during the race. After just a couple of laps they can predict the performance of each subsystem with up to 90% accuracy. And since most of these subsystems can be tuned during the race, engineers pit crews and drivers can proactively make minute adjustments throughout the race as the car and conditions change." Further details

can be found on Laney's blog (see www.blogs.gartner.com/doug-laney/the-indy-500-big-race-bigger-data/), and it is obvious that the situation for Formula 1 cars or the NASCAR series is similar. Exploitation of sensor data from cars is, however, no longer limited to race cars, as will be seen shortly.

Smart Homes

An area that will finally experience wider dissemination with the application of big data is *home automation*, a field that has been under development for more than 15 years now, but which so far has not taken off on a large scale. With big data tools there now exists the technical knowhow to process data from air conditioning and heating units, lighting, windows, doors, audio and video equipment, even household appliances such as washers, dryers, and refrigerators in conjunction with personal information from the people living in the house, in order to create living conditions optimally adapted to the specific needs of the residents. Figure 3.4 indicates the many application areas for smart home concepts, and there are numerous companies are already active in one or more of these areas. The important point is that data is produced everywhere and is ideally combined in a way that can lead to useful automation and subsequently enhanced convenience for the residents.

Healthcare

In home automation the intention is to improve the living conditions of people in their homes and a by-product of this is a greater likelihood of being able to live independently for longer, even into old age. Creating better living conditions is also part of the broad domain of *healthcare*, which is increasingly supported by or based upon data gathered about a person's medical condition, daily activity, nutrition as well as other input, e.g., from drug manufacturers, and its appropriate processing. A major trigger here has been the arrival of personal tracking devices or "wearables" such as the Fitbit One, Flex or Alta, the Nike+ Fuelband, the Jawbone Up or smart watches with health monitors, which typically communicate with a local device such as a smartphone over Bluetooth and with a website over the Internet.

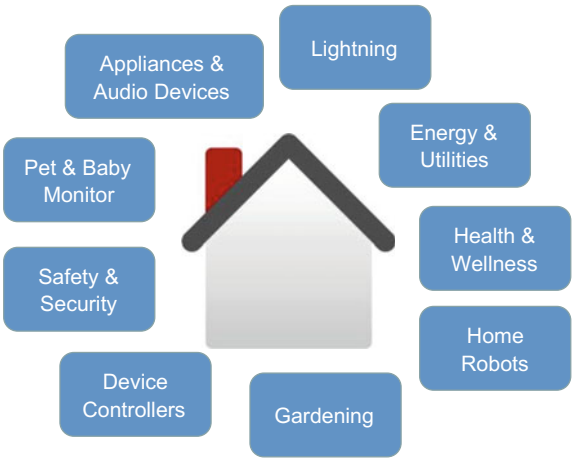

Fig. 3.4 Smart home applications areas

Commonly you need to sign up to one of the providers' sites and can then inspect your personal statistics. Activities are recorded on a daily basis; the user can set goals, monitor whether they have been achieved, and even compared himself or herself with friends who are using the same type of device. Such devices and associated activities are supported by a range of gamification-like features including rewards and leaderboards.

While this data may be beneficial for its users, for example to watch the progress of a personal diet or to find out how the personal shape has improved (or deteriorated) over time, there is a host of other people and institutions interested in that data, first and foremost your doctor as well as your health insurance provider. While a personal doctor can so far diagnose a patient only based on the data from a recent check-up, combined with the records which the doctor may keep about this patient or which may be available from previous consultations, he or she can now integrate this with fitness data which the patient provides himself. It will thus become easier, for example, to relate a heart condition to a lack of exercise, and the patient will even be enabled to control recovery him/herself. On a larger scale, it has been predicted by IBM Research (see http://www.research.ibm.com/cognitive-computing/machine-learning-applications/targeted-cancer-therapy.shtml) that "in five years, doctors will routinely use your DNA to keep you well." The healthcare area will particularly boom soon due to an availability of personal sequence or genome data and an increasing understanding of which of its portions (i.e., genes) are responsible for what disease or defect.

Similar to personal doctors, health insurance companies will likely ask for the availability of such personal tracking data soon, and they will typically be able to execute an even wider integration of data than a general practitioner, thanks to the digitization of research and test results, insurance claims, or home monitors, all of which deliver data in addition to what the general practitioner (GP) already can acquire. It can also be expected that it won't be long until the insurance premium a person has to pay for such an insurance will reflect the person's willingness to do health monitoring or to grant access to personal data.

Property Insurance

Developments analogous to healthcare can already be seen for car insurance, for example in the USA or in the UK, where some companies already reduce the premium a car owner has to pay if the latter is willing to plug a small device into the on-board diagnostics (OBD) port of their car, through which the insurer can permanently monitor how the driver is behaving when on the road. The OBD port allows access to sensor readings from a variety of devices that are built into a car. Insurer Allstate is marketing its Drivewise device as follows[6]: "Drivewise is a way for smart drivers to get rewarded for driving safely every day. Each time you take a drive, it collects feedback on driving behaviors including hard braking, high speed and when you're behind the wheel. The safer you drive, the more you can earn! Drivewise will never raise your rates. The focus of Drivewise is to give you

[6] www.allstate.com/drive-wise.aspx

feedback that can only help your driving, and your rates." Competitor Progressive is marketing its Snapshot device as follows[7]: "The fair way to pay for car insurance. It just makes sense—insurance should be based partly on how you actually drive, rather than just on traditional factors like where you live and what kind of car you have. That's what Snapshot is all about. Your safe driving habits can help you save on car insurance."

Connected Cars

One important point in these developments, be it healthcare, homes, or cars, is that we are observing more and more cases where devices talk directly to each other, instead of just recording, say, a measurement on a website and have it ready for human or algorithmic inspection. For example, in the car insurance example, the plug in a car is not talking to the driver, but to a machine on the insurer's site which can immediately use the transmitted data for premium calculations. Ultimately, (small or big) machines talk to other (small or big) machines, potentially through various stages or intermediate machines, and ultimately come up with a decision on an issue that impacts human beings. This is a typical example of the Internet of Things (IoT) and of a cyber-physical system which we will say more about in Chap. 6.

Take connected or self-driving cars. The idea, around as a vision since the 1950s,[8] is that a car be able to self-navigate along a roadmap and while doing so observe exceptional conditions (such as construction sites) and communicate with other cars as well as the road itself. Similar to race cars, autonomous cars carry sensor and camera technology, so that they can monitor their distance from other cars and their surroundings, adapt their speed appropriately, and can recognize obstacles or oncoming traffic. Ultimately, they will even be able to predict technical malfunctions or even breakdowns, and will then communicate to the nearest garage to arrange repair. When the car reaches the garage, alternative transportation will already be ready to take over the passengers; this was originally envisioned by HP Labs in California within their CoolTown project.[9]

On the downside, preliminary experience with self-driving cars as gathered by companies like Tesla, Jeep, or Volvo shows that there is still a lot to be done before the vision of totally self-driving cars will become a reality. Autonomous cars need to be connected to the Internet in order to be able to communicate with a remote server that can evaluate, for example, distance measures or images in real-time. These connections could become subject to hacking, or the decision that the car makes in response to a server communication may simply be wrong. Worse, the problem of making a "correct" decision when there are two alternatives, both of which imply casualties but in differing amounts, has been shown by Englert et al. (2014) to be closely related to the famous halting problem for Turing machines and is hence undecidable; see also Achenbach (2015). The halting problem states that there is no algorithm which can decide given a program and an arbitrary input to

[7] www.progressive.com/auto/snapshot/

[8] See www.youtube.com/watch?v=F2iRDYnzwtk

[9] See www.youtube.com/watch?v=U2AkkuIVV-I

that program, whether the program will halt for that input. While originally stated for the formal computational model of Turing machines, it generalizes to programs written in arbitrary, yet "Turing-complete" languages. Undecidability means that there is no algorithm, of whatever complexity, that can solve the problem at hand. Hence, connected cars have a built-in problem, which cannot even be solved by algorithm and even less by an ethics committee!

Smart Cities

While connected or autonomous cars are still in an early stage of development, other developments regarding transportation (or more generally, "becoming smart") are more advanced. We mention two examples next that are representatives of the positive effects (big) data can have for the customer or more generally the individual. The first example deals with Milton Keynes, a town in Buckinghamshire, England. Milton Keynes, or MK for short, has set out to become a role model for smart cities, taking a holistic view of transportation, energy and water management, enterprises and citizens. On its website www.mksmart.org/ the city originally stated: "Milton Keynes is one of the fastest growing cities in the UK and a great economic success story. However, the challenge of supporting sustainable growth without exceeding the capacity of the infrastructure, and whilst meeting key carbon reduction targets, is a major one. MK:Smart is a large collaborative initiative, partly funded by HEFCE (the Higher Education Funding Council for England) and led by The Open University, which is developing innovative solutions to support economic growth in Milton Keynes. Central to the project is the creation of a state-of-the-art 'MK Data Hub' which supports the acquisition and management of vast amounts of data relevant to city systems from a variety of data sources. These include data about energy and water consumption, transport data, data acquired through satellite technology, social and economic datasets, and crowdsourced data from social media or specialized apps."

One of these apps concerns transportation, where the goal is to provide cloud-enabled mobility (CEM) for everybody. The idea is "to connect users with information and other cloud-based services (e.g., booking and billing systems) in such a way as to reduce travel frustrations and congestion, and also allow users to make spontaneous public transport decisions."[10] Central to this is the MotionMap "that continuously describes the real-time movements of people and vehicles across the city. It will include embedded timetables, car parking, bus and cycleway information and estimates of congestion and crowd density in different parts of the city." Hence, users of the motion can, for example, decide to switch from car to bus while on their way, or switch from one type of public transport to another, since there are enabled to "see" what is happening on their path. It should not come as a surprise that all of this is based on intensive data analysis utilizing several of the techniques and technologies we have mentioned, including recent hardware developments, social media analytics, cloud computing, and recommendations.

The second example is Urban Engines, a Silicon Valley startup whose original mission was to improve "urban mobility—saving you and everyone else time in

[10] www.mksmart.org/transport/

transit—by using information from the Internet of Moving Things." The latter refers to transit systems like metros and buses, delivery services, or on-demand fleets which move through a city, thereby generating huge amounts of data. Urban Engines collected and analyzed that data in such a way that people (or companies) could understand better how traffic flows change during the day; ideally, this knowledge can be exploited to optimize a personal transportation schedule, for example by learning that a bus will be late due to a traffic jam and suggesting the user to switch to the underground, and ultimately save time. The approach also works for transportation services themselves, and Urban Engines' software has been deployed by Singapore's Urban Redevelopment Authority (URA).[11] The important point here lies in the combination of data from a variety of sources, and in analyzing this data jointly in order to identify, for example, commuter flows to and from home and work locations. Urban Engines was acquired by Google in September 2016 to become part of the Google Maps team.

Other Use Cases

Other areas that are already big on an exploitation of the massive amounts of data that can be collected include market research (enabling the customer journey we mentioned above) or the entertainment industry. Disney Parks & Resorts has developed the MyMagic+ system (see www.disneyworld.disney.go.com/faq/bands-cards/understanding-magic-band/), which through the *My Disney Experience* web site and the corresponding mobile app can deliver up-to-date information on current offerings to prospective guests planning a trip to one of the Disney parks. Disney's MagicBand can be used by the guest as a room key, a ticket for the theme park, access to FastPass+ selection, or to make a purchase. Participating visitors can skip queues, reserve attractions in advance and later change them via their smartphone, and they will be greeted by Disney characters by their name. The system behind MagicBand collects data about the visitor, his or her current location, purchase history, and which attractions have been visited.

Finally, we mention that social media sites or search engines also intensively analyze the data that they can get a hold of. Indeed, Twitter regularly analyzes the tweets its users are generating, for example to identify and compare user groups, to analyze user habits, or to perform sentiment analyses on the text of tweets. Similarly, Facebook is interested in the number of "Likes" a page gets over time and keeps a counter for recommended URLs, in order to make sure it takes less than 30 second from a click to an update of the respective counter. Google performs text clustering in Google News and ties to show similar news next to each other; moreover, they classify e-mails in Gmail, and perform various other analytics tasks, e.g., in connection with their AdWords business. We will look "behind the curtain" for some of these applications later, in order to give the reader an idea of the techniques employed in these areas.

[11] www.ura.gov.sg/uol/

3.3 Mobile Commerce and Social Commerce

In Chap. 1, we presented the key mobile technologies used by businesses today. We also introduced the concept of socialization and presented a number of applications of social media technologies. At the beginning of this chapter, we briefly summarized the commercialization of the web and e-commerce. In this section, we continue our discussion of e-commerce, but specifically look at how mobile technologies and social media are increasingly becoming the platforms of choice for both customers and retailers wishing to buy and sell products and services on the Web.

According to emarketer.com (2014), it is forecast that by the end of 2017 over two billion mobile phone or tablet users will make some sort of mobile commerce transaction. Further predictions suggest that by the end of 2018, some 27% of all US retail e-commerce sales will be carried out using mobile devices with an anticipated value of some $US 133 billion. It would seem that these predictions may have indeed been a little conservative, as the 2015 Internet Retailer 2016 Mobile 500 study found that by 2015, US mobile sales had already exceeded 29%. What these and many other similar studies clearly demonstrate is that mobile commerce will eventually, and perhaps quite soon, overtake fixed-line e-commerce.

3.3.1 Applications of Mobile Commerce

Mobile commerce can be defined as any business activity conducted over a wireless telecommunications network or from a mobile device. Specifically, in most cases it can more simply be defined as the buying and selling of goods and services through wireless handheld devices. Service-based mobile transactions can include those involving the likes of entertainment such as online gaming, gambling, and content consumption (e.g., from Netflix, Amazon Prime, or iTunes). They can also include transactions that result from a user viewing advertisements on their mobile devices. Another key source of mobile commerce revenue is from web-based mobile communication. With increasing ubiquity of low-cost wireless networking (Wi-Fi and Cellular), mobile users are moving away from more traditional, fixed-line forms of electronic communication to greater use of mobile-only communication applications such as WhatsApp, Viber, or Snapchat, as well as mobile enabled applications such as Facebook Messenger, FaceTime, or Skype. All of these applications are offered free to use, albeit with limited functionality, funded through the ever-increasing presence of online advertising. WhatsApp, purchased by Facebook in 2014 for an estimated $US19 billion, is an exception and, to-date, has resisted the opportunity of advertising, instead recognizing the value of over one billion registered users and instead encouraging users to sign up to Facebook where they can contribute to its extremely successful advertising-based funding model.

3.3.2 Attributes of Mobile Commerce

While we mention ubiquity as being a key enabler of mobile commerce, the rapid growth of mobile commerce that has been observed around the world cannot be attributed to any singular factor. Indeed, there are a number of reasons behind what cannot only be described as a highly disruptive innovation. The following attributes are collectively responsible for this growth.

- Ubiquity: widespread, uninterrupted network access enables easier information access in real-time.
- Convenience: We now have access to highly sophisticated "computers in our pockets" that store data, provide access to the Internet and intranets, offer intuitive touch screens, high definition video and sound, with ever increasingly acceptable battery life.
- Instant connectivity: It seems a distant memory nowadays the process we had to follow in order to connect to the Internet, and that was via a wired connection. Today we are provided with easy and quick connection to the Web, as other mobile enabled devices through the likes of Bluetooth or NFC.
- Personalization: What we see on our devices has been specifically tailored to our own specific needs and wants. This, based on highly sophisticated Big Data analytics, collaborative filtering, and recommender systems provides a mobile service that appears to have been developed by somebody who know us intimately.
- Localization of products and services: This more recent development taking advantage of inbuilt GPS technologies and, in the future, beacon technologies, is concerned with knowing where users are located at any given time and matching services to them.

3.3.3 User Barriers of Mobile Commerce

While mobile commerce appears, on the face of it, to offer consumers a shopping/entertainment experience comparable to that of fixed-line e-commerce but with the ability to engage from any location, there are some consumer challenges that developers and service providers alike need to be aware of. The three main issues that concern smartphone users and prevent them from using their devices to engage in m-commerce revolve around safety and security, connectivity, and screen size.

In terms of safety and security, users fear that their devices will be attacked by viruses, resulting in the theft of personal data. Indeed customers regularly report that they feel safer, when engaging in web-based transactions, when they are at home using fixed-wire communications. The home setting gives the buyer familiarity and the software and technology sitting on a desk at home feels more secure and protective. This perception, however, is not necessarily supported by the mobile industry. For example, McCaskill (2015) quotes leading specialist security company Kaspersky and top mobile payments provider Zapp as suggesting that

mobile security is actually probably safer than that of your average laptop. The reason for this is still a little unclear, but what has been observed is that hackers are still predominately focused on fixed-line online transactions, not so much mobile.

While mobile connectivity continues to improve, at least within urban areas within development nations, there continues to be concerns of the quality of mobile connections, especially when financial transactions are involved. Users remain concerned over slow or unstable connections and in particular worry that they may be cut off in the middle of an e-commerce transaction. In the vast majority of cases however, mobile application providers have technical solutions for such situations. Increasing network speed, availability and reliability will continue to reduce the likelihood of such occurrences.

One mobile commerce challenge that is less easy to solve relates to the small screen size that inherently characterizes mobile devices. While on the one hand, users demand the convenience that goes with small-sized devices, on the other hand they want to be able to view in detail the products and services they are purchasing. Unless a buyer is familiar with a product or service, or the product's appearance does not matter, users report being hesitant to buy an item on a smartphone. This has led to many technical and design developments to try to maximize the space available on a mobile screen and provide a user experience that is both intuitive and maximizes the capabilities of the screen for which the user is viewing product and service information. Further technological developments around foldable and expandable screens will continue to reduce these user concerns.

3.3.4 Social Commerce

While mobile technologies are disrupting the "where" in terms of e-commerce transactions, social media is having a similar impact on "how" we are conducting these transactions. Social commerce, sometimes abbreviated as "s-ecommerce", is a term often used to describe new online retail models or marketing strategies that incorporate established social networks and/or peer-to-peer communication to drive sales. Cohen (2011) more succinctly states that social commerce "is the evolution and maturation of social media meets shopping." More specifically, social commerce, is, as represented in Fig. 3.5, the intersection between e-commerce and social media.

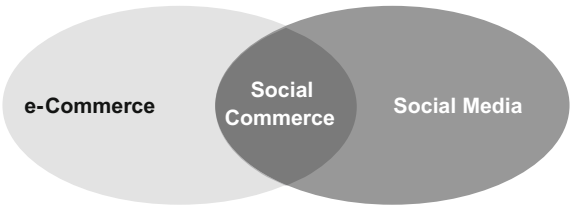

Fig. 3.5 Defining social commerce

It is helpful when considering the role of social commerce to consider an analogy, based around two central items—cash registers and water coolers. If we consider putting water coolers next to cash registers, that represents helping people connect where they buy by adding and linking social media tools and content to e-commerce sites, e.g., Amazon—rate and review products. Putting cash registers next to water coolers is the analogy for helping people buy where they connect by embedding social media stores and storefronts to popular social media platforms, e.g., Best Buy's storefront in Facebook.

In the past, business use of social media has been restricted to a use of, typically, Facebook for marketing and promotional purposes. However, with greater functionality being introduced in such applications, social media is increasingly being used to improve customer retention and build brand loyalty, contributing to the customer journey (see Fig. 3.1). To this end companies are learning to identify opportunities to maintain customer engagement strategies after an initial purchase. While social commerce is still a small percentage of online retailing, its growth rate exceeds all others, with Internet Retailer's Social Media 500 (2015) reporting year on year growth between 2013 and 2014 of 26%. That compares with the roughly 16% growth for the overall e-commerce market in the US.

It is instrumental to consider why such growth has occurred, and to fully understand this, the business and customer perspectives need to be separately considered. From the business perspective, social commerce aids with marketing monetization, i.e., helps marketers monetize and measure campaigns. It also contributes to e-commerce sales optimization by improving conversion rates and increasing average order value. Finally it is used by businesses for creating new revenue streams by curating and extracting value from social media content. From the consumers perspectives, trust, utility, and fun characterize what appeals. Perceived trust increases because social media content increases the "source credibility" of sales and marketing messages, making them more believable, persuasive, and trustworthy. In terms of utility, by putting social commerce tools at the disposition of customers, brand, businesses, and retailers can enhance the online customer experience. Finally, probably most importantly, social commerce brings back the fun in online shopping. By contrast, early e-commerce was a solitary experience typified by people interacting with software. Social commerce helps make commerce social again and can enhance the entire customer journey as discussed in Sect. 3.1.

3.3.5 Dimensions and Models of Social Commerce

There are six dimensions of social commerce (Marsden 2009):

1. Social Shopping: allows customers to share their online shopping experience with others (synchronous shopping); adds emotion/feeling to the experience; enables real-time recommendations.
2. Rating & Reviews: provision of independent third-party evaluation of a product or service review, with user contributions encouraged.

3. Recommendations & Referrals: provides a mechanism to promote recommendations and referrals within social networks, providing gamified rewards for referrers; integrated in social shopping portals; use of syndication tools via various social media platforms to share recommendations with friends, fans, and followers.
4. Forums & Communities: used to connect customers and businesses to each other in a moderated and curated environment.
5. SMO (Social Media Optimization): used to promote and publicize websites and website content through social media.
6. Social Ads & Apps: branded content in social media in the form of paid advertisements or social applications.

There exists a number of different social commerce business models. Not all utilize social media extensively, but instead involve elements of socialization that exist within existing retailing platforms. Social commerce business models, as defined by Sagefrog (2013) and wpress4.me (2013) include:

- Peer-to-peer sales platforms (eBay, Etsy, Amazon Marketplace): community-based marketplaces, or bazaars, where individuals communicate and sell directly to other individuals.
- Social network-driven sales (Facebook, Pinterest, Twitter): sales driven by referrals from established social networks, or take place on the networks themselves (i.e., through a "shop" tab on Facebook).
- Group buying (Groupon, LivingSocial): products and services offered at a reduced rate if enough buyers agree to make the purchase.
- Peer recommendations (Amazon, Yelp, JustBoughtIt): sites that aggregate product or service reviews, recommend products based on others' purchasing history (i.e., "others who bought item x also bought item y," as seen on Amazon), and/or reward individuals for sharing products and purchases with friends through social networks.
- User-curated shopping (The Fancy, Lyst, Svpply): shopping-focused sites where users create and share lists of products and services for others to shop from.
- Participatory commerce (Threadless, Kickstarter, CutOnYourBias): Consumers become involved directly in the production process through voting, funding and collaboratively designing products.
- Social shopping (Motilo, Fashism, GoTryItOn): sites that attempt to replicate shopping offline with friends by including chat and forum features for exchanging advice and opinions.

Looking forward, yotpo.com identifies participatory commerce, social shopping, curated shopping, and peer recommendations as the business models that will flourish in the future.

3.4 Social Media Technology and Marketing

In Chap. 1 we discussed the impact that the Web has had on the growing importance of social networks over the years. A particular result of this has been the wide emergence of social networking sites as well as blogs (e.g., Wordpress), microblogging (e.g., Twitter), or wikis (e.g., Wikipedia). Readers interested in the impact these media nowadays have in general should consult sites like http://www.fanpagelist.com/category/top_users/, "the social media directory of official accounts of your favorite brands, celebrities, movies, TV shows and sports teams," or http://www.ebizmba.com/articles/blogs, which shows the top 15 most popular blogs on a monthly basis. We will focus here on specifically at the business impact of social media, how companies might use them, and how they could analyze the impact of their use.

3.4.1 Social Media and Business

Social media is no longer just confined to public usage by private people. Indeed, many companies have discovered social media for their internal communication, and nowadays use them intensively. Examples of tools available for this purpose include Socialtext (see http://www.socialtext.com/, an "integrated suite of web-based social software applications includes microblogging, user profile, directories, groups, personal dashboards using OpenSocial widgets, shared spreadsheet, wiki, and weblog collaboration tools, and mobile apps"), Atlassian Confluence (see http://www.confluence.atlassian.com/, a type of team collaboration software), Asana (asana.com, "the easiest way for teams to track their work—and get results"), Slack (slack.com, "messaging for teams"), or Starmind (see http://www.starmind.com/). The effects social media can have on an organization have been nicely summarized by Kietzmann et al. (2011) in the "honeycomb" of social media, which separates between social media functionality and implications of that functionality as follows:

Social Media Functionality:

- Presence: the extent to which users know if colleagues are available.
- Sharing: the extent to which users exchange, distribute and receive content.
- Relationships: the extent to which users relate to each other.
- Identity: the extent to which users reveal themselves.
- Conversations: the extent to which users communicate with each other.
- Reputation: the extent to which users know the social standing of others and associated content.
- Groups: the extent to which users are ordered or form communities.

Implications of the Functionality:

- Presence: creating and managing the reality, intimacy and immediacy of the context.

- Sharing: content management system and social graph.
- Relationships: managing the structural and flow properties in a network of relationships.
- Identity: data privacy controls, and tools for user self-promotion.
- Conversations: conversation velocity, and the risks of starting and joining.
- Reputation: monitoring the strength, passion, sentiment, and reach of users and brands.
- Groups: membership rules and protocols.

Clearly, not all the functionality mentioned here is commonly available in a single tool, and not everything is desirable or appropriate in every enterprise context, but the important point is that organizations have started to recognize the benefits achievable with social media and are now exploiting them widely.

An early study of social media impact on businesses was performed by Andriole (2010), where he posed questions to managers and executives such as the following: What good is Web 2.0 technology to your company? What problems might Web 2.0 technology solve? How can we use the technology to save or make money? What are the best ways to exploit the technology without complicating existing infrastructures and architectures? Andriole's study produced a number of important findings which have, since then, been supported by observed practice:

- Web 2.0 technologies can help improve collaboration and communication within most companies.
- These technologies should be assessed to determine real impact, and a number of assessment techniques, including interviews, observations, and surveys, can be used to measure impact over time across multiple business areas.
- These technologies can help improve collaboration and communication across multiple vertical industries, though many companies are cautious about deploying them.

Although this may change over time, it is interesting to note that many companies have recognized the convenience and benefits these media can offer both internally and in the interaction with their customers. Indeed, companies often allow their customers today to approach them through a variety of channels, including voice (telephone, VoIP), social networks, e-mail, classical mail, private messages, or chat and consequently employ what is a called a "multi-channel strategy" in their customer relationship management. On the other hand, the saying that "a fool with a tool is still a fool" is still valid; people's mindset must be such that they are willing to engage in all this.

3.4.2 Social Networks as Graphs

As it is our goal in this chapter to give the reader an impression of how to approach an analysis of social media, or how enterprises learn about their customers via

social media, we now take a look at a typical problem in the context of social networking, that of *determining communities*. Intuitively, communities are groups of people that are related by a common interest or purpose and that often interact regarding this interest; they also develop a sense of togetherness. Communities, once found, can often be addressed as a whole, in order to support their purpose or simply as subjects for conducting business. Communities can form for a variety of purposes and goals, e.g., people with the same disease, people forming a shopping community, people with the same hobby, the alumni of a school or university, a sports club, or the fans of a particular type of music; notice that communities often overlap, i.e., individual members often belong to more than one community.

This last remark is already a hint to a technical challenge: Communities can easily be visualized as graphs, where nodes represent individual members and edges a relationship (e.g., "friend") between two members. So one might expect that classical graph algorithms are applicable, for example for determining weakly or strongly connected components. The catch is that graph algorithms tend to determine *disjoint* subsets of the set of nodes, so that no node can be in more than one of them. This is counterintuitive when applied to overlapping communities, which is why different approaches are needed, one of which is described next.

As a running example, we consider the graph shown in Fig. 3.6, which shows a very small social network. The interpretation of this graph is that there is a collection of participating entities, e.g., individuals which form the nodes of the graph and which in our example are named A … F. Moreover, there is at least one relationship between entities of the network, which could be absolute or with a degree, i.e., there could be different kinds of relationships between individuals, but these are ignored here. As a consequence, we can consider *undirected* graphs, where every relationship is symmetric, i.e., if X is a "friend" of Y, then Y is also a "friend" of X. Finally, there is an assumption of locality, i.e., relationships tend to cluster, e.g., if X is related to Y and Z, then Y and Z are probably also related. However, notice that relationships are not necessarily transitive: If X is related to Y and Y is related to Z, then X is not necessarily related to Z as well.

Obviously, real-world social networks are considerably more complex than our tiny example. Many examples in that regard can be found on the Web; for instance,

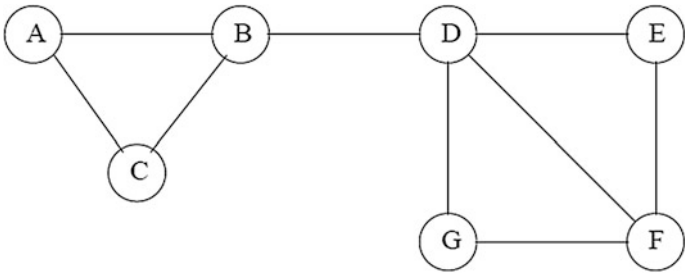

Fig. 3.6 Example of a (tiny) social network graph

Internet user John M. Baker shows his LinkedIn connections (as of 2011) at http://www.etechsuccess2.blogspot.de/2011/01/my-social-network.html.

From a formal point of view, a social network is a graph G = (V, E) with a set V of vertices (nodes) and a set E of edges that is a subset of V × V. Key problems in social network analysis are the following:

- Centrality: To what degree is a given node "central" to the network, or how important is a node in the network?
- Link prediction: Which edges currently not in the network or graph are most likely to be added at some point?
- Community detection: How can the nodes in the network be clustered into "natural" or "useful" groups?
- Information diffusion: How does information spread or diffuse over the network?

We will here just look at the problem of *detecting communities*. Informally, a community is a subset C of V such that there are many edges between the nodes in C (and considerably less between two different such subsets). When looking at Fig. 3.6, we would intuitively say that there may be two communities, one comprising nodes A, B, and C, and one comprising the remaining four nodes. In other words, there is an edge, (B, D), which is a kind of "bridge" between these communities, or which is just "between" them. In the next section, we will demonstrate how the determination of the "betweenness" of edges can be seen as a key to community detection.

3.4.3 Processing Social Graphs

We will now describe an algorithmic approach to finding communities in a social network. Although we perceive such a network as a graph, traditional graph algorithms, such as those for finding strongly or weakly connected components, will not work, as we mentioned above. An additional complication, besides the fact that classical algorithms avoid overlaps, is that they are typically based on a measure for determining distances between nodes. There is, however, a catch in social networks, namely the fact that in undirected graphs where the edges represent "friend" relationships, there is no suitable way to define a distance measure. We could say, for example, that nodes are *close* if there is an edge between them (and *distant* if not): For x, y ∈ V, distance d(x, y) is 0 if (x, y) is in E and 1 otherwise. However, now consider the case that edges (x, y) and (y, z) are present, while (x, z) is not. Then the triangle inequality, whose validity is often a basic assumption in graph algorithms, would require that $d(x, z) = 1 \leq d(x, y) + d(y, z) = 0 + 0 = 0$, a contradiction! So the triangle inequality does not hold, and we need a different algorithmic approach (we will later see, for example in connection with online advertising that also in other Web applications traditional algorithmic approaches are now longer usable).

Following the example shown in Fig. 3.6, we are now interested in finding edges that are least likely to be inside a community. We will make this precise using the notion of "betweenness" of an edge (x, y) ∈ E, defined as the number of pairs of nodes u, v such that (x, y) lies on the shortest path between u and v. For example, in Fig. 3.6 edge (B, D) has the highest betweenness of any edge in this graph, since it appears on *every* shortest path between any of the nodes A, B, C to any of the nodes D, E, F, G; all these 12 paths go through (B, D); hence betweenness (B, D) = 3 × 4 = 12.

The intuition behind the notion of (edge) betweenness is to look at the strengths of weak ties, and the higher that strength, the weaker the tie. This is similar to playing golf, where a high score is also bad; high betweenness of (a, b) suggests that (a, b) runs between two different communities, yet a and b do not belong to the same community.

We mention that there is a related centrality notion, *node betweenness*, which considers individual nodes instead of edges. It indicates how "central" a node is in a network and is again measured by the number of shortest paths from all vertices to all others that pass through that node. A node with high (node) betweenness centrality has a large "influence" on things (messages, opinions) that pass through the network, under the assumption that passing is based on shortest paths. Both concepts have many applications, including computer and communication sciences, biology, transport and scientific cooperation.

We next present an algorithm originally proposed by Girwan and Newman (2002), hereafter abbreviated as GNA. It focuses on edge betweenness and detects communities by progressively removing edges from the original network, in such a way that the connected components of the remaining network are the communities (which may have smaller communities nested inside). GNA focuses on edges that are most likely "between" communities, and essentially proceeds in three steps as follows:

1. First, the betweenness of all existing edges in the given graph is calculated, by considering each node X in turn, determining the number of shortest paths from X to other nodes, and using that number for assigning a partial betweenness to the edges adjacent to X. When all nodes have been considered, add the betweenness values determined for each edge and divide by 2 (since for each edge, both its endpoints have been considered).
2. The edge with the highest betweenness is removed (if multiple edges have the same highest betweenness, remove them all). The graph may thus split into several disjoint components; if so, we have found some communities already.
3. Now make the second highest betweenness the highest and repeat Step 2 and until the graph is broken into a suitable number of connected components, which form the communities.

In the example of Fig. 3.6, the calculation of edge betweenness will yield the result shown in Fig. 3.7; as mentioned before, the edge with the highest betweenness is (B, D). If we remove this edge from the graph, we obtain two communities as expected, one with nodes A, B, C and one with nodes D, E, F, G.

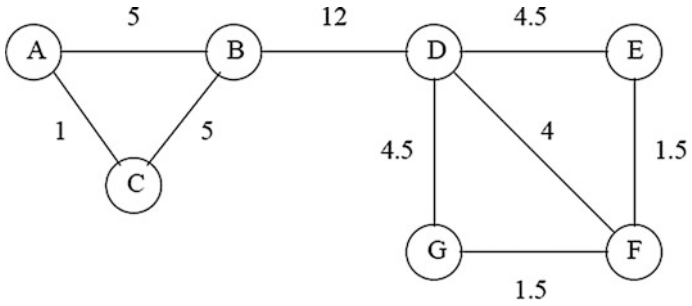

Fig. 3.7 Graph from Fig. 3.6 after first betweenness calculation

We could then consider edges (A, B) and (B, C) in the first component as well as edges (D, E) and (D, G) in the second, whose removal would yield smaller communities; whether these are meaningful, however, would have to be answered from an application point of view.

We are not discussing GNA in further detail, but mention that the algorithm essentially needs to consider all the shortest paths between all pairs of nodes, which is computationally expensive and can hence become an obstacle when determining communities in networks with thousands or millions of nodes. GNA avoids this by using a breadth-first search approach, which by counting the number of shortest paths from a node to every other node determines the "flow" values for each edge. This method has a complexity that is proportional to $n \times e$ (i.e., linear in the size of the input), where n is the number of nodes and e is the number of edges in the given graph, which essentially means that it can be solved efficiently. Details can be found, for example, in Leskovec et al. (2014).

To conclude this section, we note that there are many other techniques for analyzing (social) networks. For example, in a site where users can post pictures and tag them with keywords, the interest may be in "similar" pictures, where similarity could either be based on a use of similar tags, or on similar picture content, or both.

More generally and according to the position2 blog,[12] there are four different types of tools that enterprises need for analyzing social networks:

1. "**Listening Tools**: A brand cannot afford to be ignorant about what's being said about it on any major social platform. Social media listening can be your digital eyes and ears. They sift through all the chatter and analyze it for positive and negative comments. Depending on their complexity and features, they can give you alerts on where your brand is featured or direct your attention to negative comments and potential trouble creators.
2. **Reach Tools**: Every brand wants to maximize its reach on social media. With the variety of social media platforms today, this is becoming an increasingly

[12] www.blogs.position2.com/four-types-social-media-analytical-tools-need

tough task. Each platform has different formats—Slideshare can carry a 100 MB presentation but Twitter restricts at 140 character messages. The growth and diversity of social media offerings makes manual posting across social media platforms tougher with each passing day.

3. **Depth Tools**: Some products are high involvement products and choosing the right product means a lot to the buyer. It could be a camera for an amateur photographer or a financial software package for a small business owner. The stakes are high for the buyer either emotionally, financially or in terms of effort and impact.
4. **Relationship Tools**: Social relationship tools are useful for publishing content on social sites. These tools offer scheduling capabilities, which ensure an enduring online presence. This helps the brand stay in touch regularly with consumers instead of making sporadic appearances."

Important tools are henceforth tools that perform sentiment analysis, cluster analysis (such as GNA) as well as other forms of analytics that ideally result in perceptions that not only allow us to study the relationship between a company and a customer as it has been in the past, but that ideally allows to foresee how to improve (or reestablish) it in the future.

3.5 Online Advertising

We mentioned in Chap. 1 that advertising on the Web has become one of the most prominent Internet business models. It begun with simple *banners* that could be placed on other Web sites. Since then it has emerged as one of the major ways to make money on the Web, which according to Battelle (2005) is due to Bill Gross and his invention of GoTo, a service that became famous for being among the first to differentiate Web traffic. Indeed, what Gross quickly realized was that non-targeted advertising on the Web was largely irrelevant and of little value to the advertiser as long as the traffic passing by any placed ad was the "wrong" traffic, i.e., from users not interested in what was being advertised. If users arrive at a site due to a spammer who has led them there, due to a bad portal classification, or due to a bad search result, they are unlikely to be interested in the products or services offered at that site. He hence started investigating the question of how to get qualified traffic to a site, i.e., traffic with a reasonable likelihood of responding to the goods or services found at a site, and then started calculating what businesses might be willing to pay for this. This gave birth to the idea that advertisement can be associated with the terms people search for and the pay-per-click tracking models we see today in this business.

Advertising has become a major business model on the Web since the arrival of Google AdSense, according to them "a fast and easy way for website publishers of all sizes to display relevant Google ads on their website's content pages and earn money" and Google AdWords, which allows businesses to "create ads and choose

keywords, which are words and phrases related to [their] business. … When people search on Google using one of [the] keywords, [the] ad may appear next to the search results." It is hence no surprise that consumers of stationary or mobile devices are constantly flooded with ads today. It is also a major source of income for social networks like Facebook, and hence is a connection to the topic of community detection we discussed in the previous section.

Before we embark on a more detailed discussion of online advertising as a business model in general and AdWords in particular, we note that advertising on the Web represents another incarnation of the long tail curve of Web applications we have seen in Chap. 1 (Fig. 3.7): Through Google AdWords and related approaches (e.g., in social networks), it has become possible not only for large companies (amounting to 20% of all companies) to place advertisements on the Web, but now the same is possible even for a small company. Through a cost-effective and highly scalable automated infrastructure provided by the index of a search engine or of a social network, online sites can offer advertising even for very limited budgets, as they may be only available for a small company. In other words, small companies do not have to set up an infrastructure for advertising (even in niche markets) themselves, they can simply rely on what others are providing and searching for on the Web.

The 20+ years of history of online advertising has seen a variety of important developments, including the following:

- **Direct placement**: Advertisers post their ads directly on a site, and do so for free or for a fee or pay a commission. Examples include eBay, craigslist, and many auto trading sites. The selection of an ad by a user can be based on parameters (e.g., make, model, or year of a car), or it can be done relative to query terms (e.g., "apartment Belmont"). Ranking of ads, i.e., the question of in which order ads should be presented, is typically tricky under this approach and may be based on such strategies like "most recent first" or similar. However, users can be shown individual ad selections.
- **Display ads**: These are banners that are placed on many sites, sometimes at fixed places (e.g., upper left corner), sometimes in the middle of text that represents a search result or an article spread over multiple Web pages, yet all users get to see the same ads. Banner ads resemble advertising in traditional media (e.g., magazines, TV); however, a big difference is that the advertiser now typically pays for impressions, not just for showing the ad. An obvious benefit is that the Web can exploit the information about its users, in order to determine which ad they should be shown; this information can be gathered from a variety of sources, including social media, email, bookmarks, time spent on a page, or search queries issued. For example, if a search engine recognizes a user (via cookies or when logged into an account with the respective provider) and can record that the user has an interest, for example, in motorsports, there is a high probability that an advertisement for car and car parts will be regularly presented

to that user. As mentioned earlier, banner ads were the initial form of advertising on the Web and have brought along a typical foundation for calculating fees, the CPM (cost per mille,[13] i.e., per thousand impressions) rate, meaning that an advertiser does not have to pay for each and every click of his ad, but only per thousand. Banner ads typically show low click-through rates and, correspondingly, a low return on investment or revenue for the respective advertiser.

- **Search advertising**: This form of advertising is based on the simple idea of creating an association between what a user is searching for on the Web and the ads shown to her or him in response to a search query. This idea was originally developed by a company called Overture, which was acquired by Yahoo! around the year 2000, and would place ads together with the results of a search query; advertisers can now bid on certain keywords, and when someone searches for one of these keywords, the ad of the highest bidder is shown. Like with banners, the advertiser is charged only if the ad is actually clicked on. The concept is (and has been) easily extended to e-mail, where a provider can analyze e-mail content, e.g., search for "important" terms in e-mail, and then select and show ads correspondingly.

Search advertising has become a primary advertising method on the Web since Google adopted it around 2002. After a number of changes it was made available to the public under the name *AdWords*. We will discuss these changes below, but in particular focus on two issues that arise when ads are shown dynamically. These are:

- How to determine which ads should be shown together with a particular search result?
- How to rank ads in case multiple ones link to a given search term?

Additional questions, not discussed here, refer to the question of how to attract views and clicks, where to place an ad on a Web page, and generally how to do better than traditional mass-media (radio, TV billboards), both from a provider's and from an advertiser's point of view.

3.5.1 A Greedy Algorithm for Matching Ads and Queries

We next look at the question of which advertisement(s) should be displayed with the results of a given query. So the situation we consider is that of a search engine which can answer search queries, and which has decided to monetize the ad business. More specifically, when a user places a search, say, on "sports car," the search engine would come back with links to Web sites on sports cars, but also show advertisements of sports car vendors and manufacturers, and vendors of associated accessories. To make this happen, the advertisers have previously stated

[13] Latin for thousands.

in one way or another that they are interested in having their ads placed near search results for "sports car." The typical way they indicate this is through an online auction, in which advertisers offer a certain premium they are willing to pay to the search engine provider as soon as their ad is clicked. We assume that every advertiser or bidder has a certain budget (e.g., for the month), and that budget cannot be exceeded.

The following simple example indicates that the letting the highest bidder win might not always be the best solution: Suppose we are faced with the following situation:

Advertiser	Bid on "Mustang" (m)	Bid on "Camaro" (c)	Budget
A	2	1	5
B	1	2	5

Thus, we only have two bidders and no one else; both still have all of their budgets, and we only display one ad per query. Next assume we receive the following sequence of search queries:

c m c m c m

The first query asks for "camaro", the next for "mustang", the next for "camaro" again, and so on. If ads are placed as follows

B A B A_ _

The search engine will get stuck after answering four queries, since B, the highest bidder on "c", does not have enough budget anymore, and similarly, A, the highest bidder on "m" does also not have enough budget anymore; thus, the overall revenue will be 8, and no more ads will be shown after the first four searches.

So what we see here is that highest bids are not always a guarantee for having ads placed, and indeed a search engine will typically keep track of how often the ads of a particular advertiser are actually clicked, in order to get a more realistic picture of potential revenues; we will come back to this point later.

We could do better than what we saw above if the search engine had an idea of the future or, in other words, would know in advance which search queries to expect. Indeed, if the entire sequence "c m c m c m" had been known in advance, the search engine could have assigned ads as follows:

B A B A A B

Now the last two ads are not from the highest bidders anymore, but the overall revenue would come to 10.

It is easy to see that this aspect of "knowing the future" can make a crucial difference. Consider the following revised example, where budgets remains the same as above, but bids go down:

Advertiser	Bid on "Mustang" (m)	Bid on "Camaro" (c)	Budget
A	1	0	5
B	1	1	5

Now assume that the search sequence is

m m m m m c c c c c

(i.e., five queries for "m" followed by five queries for "c"). Since both advertisers bid on "m" and there is no difference in their bids, the search engine might assign as follows:

B B B B B_____

But then B's budget is exhausted, and A does not bid on "c", so the revenue obtained in this way is 5. Had the search engine known what to expect after the first five queries, it could have placed ads as follows:

A A A A A B B B B B

and the total revenue had been 10.

What we see from these simple examples is that a type of algorithm is needed which does not wait for its input to be complete. Instead, there is a partial input to start with, and the algorithm needs to make a decision or produce an output as soon as the next piece of input is received. An algorithm of this kind is called an "online" algorithm, as opposed to an "offline" algorithm that only starts its processing once the input is completely available (which is the case, for example, for sorting algorithms like Quicksort or Heapsort)[14]. In the case of a search engine, an offline algorithm would collect, say, a month of search queries, then look at the bids and the budgets of its advertisers, and finally compute an assignment of ads to query results that optimizes revenue as well as the number of impressions, but obviously that search engine would not be of much use in a fast digital world.

Instead of waiting for more input, a search engine needs to employ an online algorithm that can instantly select an ad to be shown with the query result when a search query arrives, and the only information it can utilize is the advertisers' bids and budgets and information about the past, in this case how often the ad of a particular advertiser has been clicked (the click-through rate or CTR). Our second

[14]Notice that "online" does not mean in this context that it has to be done on the Internet; it only indicates incomplete input information.

example above shows why the future could help, but in this scenario knowledge about the future is not available.

The last example above also shows that typically an online algorithm will achieve results that are not as good as what could be achieved by an offline algorithm. Indeed, the result of the online algorithm indicated could only be 50% of the revenue otherwise achievable, i.e., by an optimal algorithm A. This discrepancy is measured by a coefficient called the *competitive ratio* c of the online algorithm at hand; in our case, $c < 5/10 = ½$, and it can be shown that the converse also holds and hence $c = ½$. In other words, the result that this "greedy" online algorithm can achieve cannot be guaranteed to be any better than 50% of the optimal result.

Having now established a basic understanding of what kind of algorithm is needed for solving the problem of matching advertisements to search queries (or their results), we now look at a simplified version of the matching problem itself. The simplification is that we consider bipartite graphs, i.e., graphs whose set of nodes can be divided into two disjoint subsets such that edges only connect nodes that belong to distinct subsets. For example, Fig. 3.8 shows a bipartite graph with four nodes in each node subset.

We interpret the edges in such a graph as *preferences*; if we consider ads to form the left set and queries the right one, our interest is in finding as many *matchings* as possible, i.e., subsets of the edges such that no node is an end of two or more edges. For example, {(1, c), (2, b), (3, d), (4, a)} is a matching for the graph shown (with thick lines) in Fig. 3.8; it is even a "perfect" matching in the sense that every node of the graph appears in the matching. A matching that has the most number of edges possible in a given graph is called a *maximal* matching. The matching just considered is also maximal, since no other matching could have more edges.

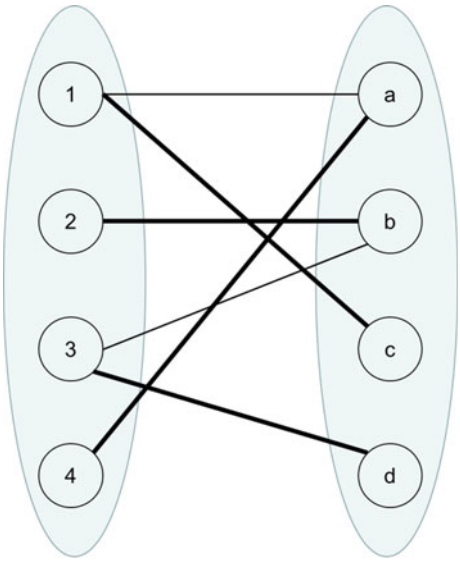

Fig. 3.8 A bipartite graph

We mention that another, possibly more intuitive interpretation of the scenario considered here is to consider the left set as boys and the right set as girls; the goal would then be to match each girl to a boy which she "likes" (or is connected to through an edge). This interpretation also indicates the online character of the scenario: Given the boys (ads), the girls (queries) arrive and need to be assigned a boy (ad) based on preference (existing bid and remaining budget). Again, if the entire bipartite graph is known in advance, the problem of finding a maximum matching is well-studied and can roughly be solved in time quadratic in the number of nodes of the graph (Hopcroft and Karp 1973).

If that is no longer the case, i.e., if we do not know the entire graph from the outset, but only the left set (i.e., the boys or ads), we need an online algorithm or a greedy matching; this works as follows: Given the set of boys, the girls arrive one after the other. Upon each arrival of a girl, her preferences (edges) are revealed, and the girl is paired with the next eligible boy. If there is none, the girl is not paired. In the graph of Fig. 3.8, this will mean that nodes 1–4 are initially given. When node a arrives, it reveals that (1, a) and (4, a) are possible choice; assume (1, a) is chosen. Next node b arrives, and (2, b) and (3, b) are possible choices; assume (2, b) is chosen. Next c arrives, and no choice is possible since node 1 is already taken. Finally, node d arrives and is paired with node 3. So we arrive at a matching consisting of three edges, which is not maximal, since we already seen earlier that a maximal matching would consist of four edges.

This result would lead us to suspect that the competitive ratio c equals ¾, but we need to consider the worst performance of the algorithm over all possible inputs. If we start again with (1, a) as above, but this is now followed by (3, b), we end up with an even worse result consisting of only two edges; hence $c < ½$, and again it can be shown that ½ is also a lower bound, so that $c = ½$. This concludes our short introduction into the essence of algorithms for matching ads to queries.

3.5.2 Search Advertising

We now consider search advertising, sometimes also called the "adwords problem" after the Google AdWords system. Informally, the problem we are faced with is the following: A stream of queries q_1, q_2, … arrives at a search engine, and several advertisers bid on each query. When a query arrives, the search engine must pick a subset of advertisers and show their ads. Not surprisingly, the goal is to maximize the revenue for the search engine! Stated slightly simpler than what Google actually does, the search advertising problem can be stated as follows. Given are:

1. A set of bids by advertisers for search queries.
2. A click-through rate (CTR) for each advertiser-query pair indicating the percentage of impressions which are actually clicked.
3. A budget for each advertiser (say, for 1 month).
4. A limit on the number of ads to be displayed with each search query.

The aim is to respond to each search query with a selection of advertisers such that the following holds:

1. The size of the selection is no larger than the limit on the number of ads per query.
2. Each advertiser indeed has a bid on the query.
3. Each advertiser has enough budget left to pay for the ad if it is clicked.

As an example, consider the following scenario of three advertisers having different bids on the same search term, but exhibiting different click-through rates and hence different expected revenues for the search engine provider:

Advertiser	Bid ($)	CTR (%)	Bid * CTR (c)
A	1.00	1	1
B	0.75	2	1.5
C	0.50	2.5	1.125

Clearly, the search engine provider is interested in maximizing its revenue, or the total value of the ads selected, where each value is calculated as bid * CTR. It is therefore understandable that he will not necessarily display the ad of the highest bidder, but those that promise the highest revenue. For example, A in the table above has the highest bid, but a low CTR, B has the highest value, and C has the highest CTR. In other words, if 1,000 queries occur for the search term in question, A will most likely be clicked 10 times and yield a value of 10 c, B will be clicked 20 times with a value of 30 c, and C will be clicked 25 times with a value of 28.13 c. So the provider will obviously be more interested in C than in A.

To make the situation somewhat more complicated, each advertiser has a limited (typically monthly) budget, which is divided by 30 to obtain a daily budget, and the search engine makes sure that no one is charged more than their (daily) budget. Moreover, the CTR of an ad is essentially unknown in advance and can only be observed and monitored over time; so typically a search engine will need to start with an assumption about the click probability of a new ad.

We next present the basic ideas underlying a greedy algorithm for search advertising; to this end, we make several simplifying assumptions:

1. There is only one ad to be shown for each query.
2. All advertisers have the same budget B.
3. All ads are equally likely to be clicked (i.e., all CTRs are the same).
4. The value of each ad is the same (=1).

Then the algorithm simply says: *For each query*, *pick any advertiser with a bid for that query*. As an example, consider two advertisers A and B such that A bids on

m, while B bids on m and c; both have a budget of 4. Similar to what we have seen earlier, for query stream

m m m m c c c c

the worst greedy choice is

B B B B _ _ _ _

with a revenue of 4, while

A A A A B B B B

would be optimal with a revenue of 8. Again the competitive ratio can be shown to be ½, but with a simple improvement called the *Balance* algorithm it can be brought up to 0.63. This improvement picks, for each query, the advertiser who bids on the query *and* has the largest unspent budget, if that applies to more than one, it picks one of them arbitrarily.

We conclude this section by briefly discussing the Google mechanism for advertisers: It is based on an ongoing auction where advertisers can submit bids on particular keywords; bids indicate the value a click would have for the advertiser when the ad is shown. Google shows a limited number of ads only with each query. Thus, while the original (Overture) idea was to simply order all ads for a given keyword, Google now decides which ads to show, as well as the order in which to show them, and as we have seen this decision is solely driven by expected revenue. That is, the click-through rate is observed for each ad, based on the history of displays of that ad. Users of the AdWords system specify a budget: the amount they were willing to pay for all clicks on their ads in a month. As we have also shown, these constraints make the problem of assigning ads to search queries significantly more complex.

We also note that online advertising is nowadays a complex business due to the number of parties involved. It is most often not just a business involving an advertiser and a Web platform provider or (ad) publisher, but there are many intermediate steps (involving intermediaries), each of which causes an application of additional fees. As can be seen from Fig. 3.9, there is an entire ecosystem behind the online advertising business today, consisting of advertising agencies, demand- as well as supply-side platform, auction markets, and at its endpoints the advertiser and publisher. The three big players distributing ads on the Web are Doubleclick by Google, Adtech, formerly AOL, and smartadserver in Europe. It is safe to assume that each transition between any two parties in Fig. 3.9 involves paying a fee, and statistics show that roughly only 50% of the amount an advertiser spends actually reaches the publisher. More information on this can be found, for example in de.slideshare.net/andrewtweed1/thomvest-advertising-technology-overview-sept-2014 or in http://www.de.slideshare.net/ksanz15/understanding-the-online-advertising-technology-landscape.

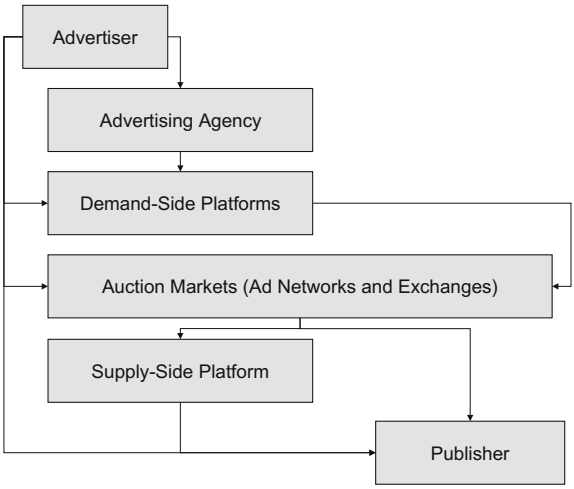

Fig. 3.9 Partial advertising technology ecosystem

We should not leave the topic of online advertising without a warning: Advertising is often misused to distribute (and execute) malware and then turns into "malvertising," a short form for "malicious advertising." The principle is simple: The malicious piece of code is hidden behind an ad that shows up when the user opens a website. When the ad is clicked, the user is not redirected to the site of the advertiser, but to an exploit landing page from which the malicious code attacks the user's computer (or device) and installs malicious software (see, for example, www.blog.malwarebytes.org/101/2015/02/what-is-malvertising/ for details).

What a user can do about this is install a so-called *ad-blocker* such as AdBlock, Adblock Plus, or uBlock Origin. This is particularly relevant to browsers that are used for any action on the Web, not just search. These malvertising networks place ads within Web sites, on mobile phones, into YouTube videos, typically in always the same way, and clicking an ad creates revenue. As we saw in Fig. 3.9, there are even marketplaces for ad space on the Web, and it is through these channels that malware finds its way into the devices or computers of end-users. Even better protection than just through an ad-blocker is to combine an ad-blocker with anti-malware software.

3.6 Recommendation

We have previously touched on the topic of recommendation: in connection with the intuition behind PageRank in Chap. 1, and in the context of big data analytics in Chap. 2. Earlier in this chapter we mentioned that it has become common in e-commerce that sites include a recommendation device. We now take a closer look at what is behind this concept.

While traditional reviews often come from a professional source (such as the publisher of a book or newspaper staff) or from private customers, online recommendations are often generated by the data mining tools that work behind the scenes; indeed, recommendations may come from other users, or they are generated from user behavior (e.g., search history, time spent on particular Web pages while browsing). Recommendation systems may look at transactional data that is collected about each and every sales transaction, but also at previous user input (such as ratings) or click paths. Ideally, it becomes possible to classify a customer's preferences and to build a profile; further recommendations can then be made on the basis of some form of *similarity* of items or categories that have been identified in or between consumers' profiles. Clearly, recommendations point to other items, where more customer reviews as well as further recommendations to more products can be found (and hopefully end in purchases).

In the context of electronic commerce, recommendations have become very popular. As an introductory example, let us briefly look at what Amazon does when a user adds a product to his or her shopping cart. Amazon then creates a special interim page where recommendations are the main pillar of the strategy, and where a mix of several strategies occurs; these are:

- Cross-selling,
- other "related" or "similar" products,
- recommended promotions,
- more generic recommendations aimed at serendipity,
- recommended products in the Amazon shopping cart.

Amazon makes the most of recommendations in the page that is displayed when you add a product to your shopping cart, seizing their opportunities to sell.

While in the context of e-commerce recommendation is typically about items or products, it can, however, also be about a number of other things in other contexts. For example, recommendation in electronic learning is about learning content, in search and navigation about links and pages, in social networks about potential new friends, or in online dating about potential dates. Irrespectively of the context, the goal is to help people (customers, users) make decisions on where to spend attention, money, time, or any combination of these. In what follows, we use "items" as generic term for what is recommended.

Figure 3.10 shows the various components of a recommender system: It incorporates two types of entity, items and users, and it takes as input ratings (if available), content data, and possibly also demographic data. Ratings can be implicit (obtained through observing user activity, including page views, purchases, or mails) or explicit, and content can be structured, unstructured (e.g., a textual evaluation), or somewhere in between. The output of a recommender system is typically a recommendation and potentially even a prediction of what a user might like next or what item could become interesting next.

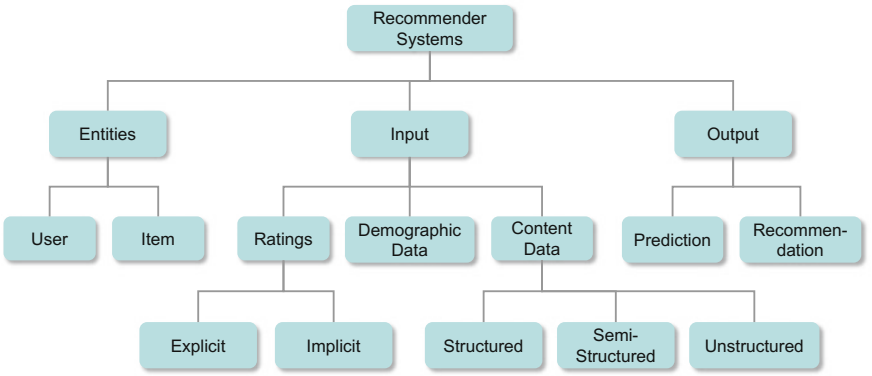

Fig. 3.10 Components of a recommender

Recommendations can come in a variety of forms, such as "Top 10," "Most Popular," or "Recent Uploads." These types of recommendations are easy to come up with, since they are primarily based on counting activities independent of a particular user. What is more relevant and, consequently, more difficult to produce are recommendations that are tailored to an individual user (like in Amazon or Netflix) and are then typically based on some form of "profile." Recommendations might also be editorial or hand-curated, and readers of recommendations should always be aware that there can also be misuse involved in what is visible.

Recommendation that takes the user into account can easily be described in an abstract way using a set S of items, a set X of customers or users, and a totally ordered set R of ratings, such "zero to five stars" or a number in the interval [0, 1]. Recommendation can then be described as a *utility function* u: $X \times S \rightarrow R$ which associates a rating with a customer-item combination. Function u can be seen as a *utility matrix* U like in the following example, where the items are movies (MI for Mission Impossible, JB for James Bond, FF for Fast & Furious), and there are four customers A, …, D; ratings are between 1 (low) and 5 (high):

	MI1	MI2	MI3	JB	FF5	FF6	FF7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

A utility matrix is commonly sparse, i.e., most entries are missing (or 0) since they are not known, since most people have not rated most items, and the goal of a recommender is to predict values for these blanks; at the least, a recommender should be able to fill in those blanks whose values are likely to be high. In reality, a utility matrix U will have many columns (items) and many rows (users; both

numbers can be in the millions or even higher), and a recommender has access to a number of attributes for both items and users in order to come up with an entry for the matrix. Notice that adding a new user would add a line to U that is initially all blank; adding a new item means adding a blank column. Even worse, a new system would start with an empty matrix and then has a "cold-start problem."

The key problems a recommender is faced with are to gather "known" ratings for U, to extrapolate unknown (high) ratings from known ones, and to evaluate the extrapolation methods employed. Clearly, the key interest when extrapolating is in high unknown ratings, since the recommender is interested in what a user likes, not what he or she dislikes. For gathering new ratings, there are again explicit as well as implicit methods, both of which have pros and cons. Explicit ratings can be obtained by asking users to rate items, which could bother people and hence lead to unreliable responses. Implicit ratings means learning from user actions, in particular from purchases that lead to high ratings, but a problem here is how to treat purchases that result in low ratings.

In the following, we will briefly discuss two major approaches to the design of recommender systems (see Fig. 3.11): *content-based* recommenders and *collaborative* filtering; hybrid recommenders as a combination of these two are also an option.

3.6.1 Content-Based Recommenders

Content-based recommenders, often abbreviated CB systems, examine properties of the items being recommended, i.e., they look at the content behind a potential recommendation. For example, if a Netflix user has watched many movies involving car chases, it makes sense to recommend movies from the "Action" genre to this user or movies with the same actors; the principle is illustrated in Fig. 3.12. For Web sites, blogs, or news entities it may make sense to recommend other sites that carry "similar" content. So the important notion here is *similarity*, and the way to decide on similarity goes through the creation of a profile; a *profile* is typically a set or a vector of features or attributes.

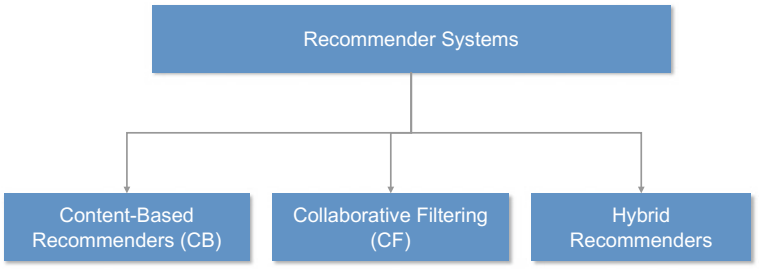

Fig. 3.11 Types of recommender systems

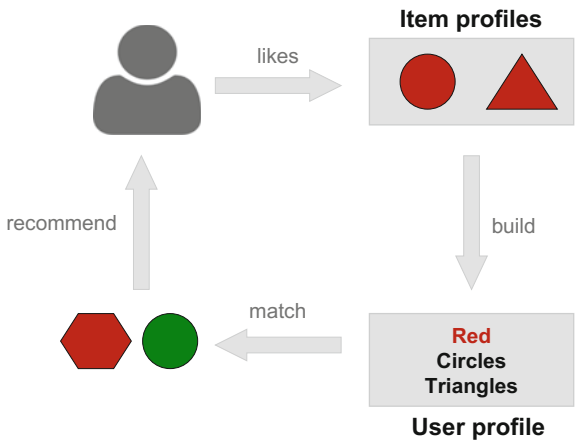

Fig. 3.12 Principle of content-based recommendation

Consider movies, which can be characterized or profiled by title, genre, main actor, director, producer, author, etc. If two movies have identical values for a "sufficient" number of attributes, they will be considered similar; if the profile P of a movie is similar to the profile P′ of another movie and user X liked P, then with some probability X will also like P′.

While profiling appears not too difficult for items like movies or cars and such, it is considerably more complicated for documents, e.g., news articles, Web pages, blog entry, or Twitter tweets. Here the question is what the "important" words in the document are. In order to determine this, techniques originally developed in the field of *Information Retrieval* are commonly applied. Obviously, the most important words in a document are not simply the ones that occur most often (i.e., words like "the" or "and" etc., so-called stop words), so another measure is needed.

A common approach to measuring word importance is based on the TF-IDF (term frequency—inverse document frequency) measure, which identifies words in a collection of documents that are useful for determining the topic of each document. Intuitively, a word has a high TF-IDF score in a document if it appears in relatively few documents, but appears in this one, and when it appears in a document it tends to appear many times.

Suppose we are given a collection of N documents. Let f_{ij} be the frequency (i.e., the number of occurrences) of term i in document j. Then the *term frequency* TF_{ij} is defined as $f_{ij}/max_k\ f_{kj}$ (f_{ij} normalized by dividing it by the maximum number of occurrences of any term in the document, in order to avoid a distortion of the result in long documents). Next let term i appear in n_i of the N given documents. Then the *inverse document frequency* IDF_i is defined as $\log_2(N/n_i)$ and measures the general importance of a term for the given document collection. Finally, the TF-IDF score for term i in document j (or its weight in j) is defined as $TF_{ij} \times IDF_i$. The terms with the highest TF-IDF score are (often) the terms that best characterize the topic of a document.

As an example, consider a document d containing 100 words wherein the word *Mustang* appears 3 times (possibly after some preprocessing, see below). Then f = 3, and the term frequency for Mustang in this single document is TF = 3/100 = 0.03. Now assume we have 2^{20} = 1,048,576 documents and the word Mustang appears in 2^{10} or 1,024 of them. Then the inverse document frequency is calculated as $\log_2(2^{20}/2^{10}) \approx 10$. Thus, the TF-IDF score for Mustang in document d is the product of these quantities: $0.03 * 10 \approx 0.3$.

With these preparations, profiling a given document by topic may proceed as follows:

1. Preprocess the document by eliminating stop words (as well as other actions, e.g., word stemming).
2. Compute the TF-IDF score for each remaining word in the document; the ones with the highest scores are the words that characterize the document.
3. Take as the features of a document the *n* words with the highest TF-IDF scores.

For example, in a document describing muscle cars, words like "Mustang," "Camaro," "Challenger," or "V8" might turn up as having the highest TF-IDF scores and would hence be considered the most important features of that document.

CB recommenders have the obvious advantage that in order to produce a recommendation for a user, no data on other users is needed, and they could even provide an explanation, why a particular item was recommended (by listing the content features that led to their decision). On the other hand, if content is not well represented by keywords, which is the case, for example, for images or music, recommendation based on the CB approach is more difficult. Also, CB recommenders cannot distinguish items represented by the same set and values of attributes, and they have difficulties to recommend items outside a user's content profile.

3.6.2 Collaborative Filtering

Recommenders based on collaborative filtering (CF) recommend items based on a notion of similarity between users or items, i.e., the items recommended to a user are those preferred by "similar" users or are simply "similar" items. It is this category that often comes in the form of "customers who bought this also bought …" Clearly, what is needed here is a similarity measure.

Before going into further detail, we briefly continue our discussion from the previous subsection. While the TF-IDF measure is intended to figure out meaning, or to classify documents by first finding the significant words they contain, and is hence appropriate for content-based recommendation, sometimes simpler methods for document comparison will do. Indeed, it is sometimes enough to look at "character-based" similarity instead of similar meanings, for example when the interest is in exact copies of a document (like in plagiarism or when looking for

mirror pages of a Web page) or in product recommendations on sites like Amazon. To this end, a variety of options exist, including the Jaccard similarity (a set-based measure) if we consider sets of terms, or the cosine similarity (a vector-space measure) if we consider vectors of terms. Suppose we have two sets of words as follows (representing, for example, distinct books that have often been bought by the same customers):

$$\begin{aligned} X &= \{\text{Mustang}, \text{Camaro}, \text{Challenger}\} \\ Y &= \{\text{Mustang}, \text{Camaro}, \text{Challenger}, \text{Veyron}, \\ &\quad \text{Regera}, \text{Miura}, 911, \text{Pantera}\} \end{aligned}$$

Then the Jaccard similarity of these sets is determined by looking at the relative size of their intersection, and hence given by

$$\text{sim}_J(X, Y) = |X \cap Y|/|X \text{ U } Y| = 3/8 = 0.375$$

Instead of using features or attributes of items to determine their similarity, recommenders in the collaborative filtering category maintain a database of many users' ratings of (a variety of) items. For a given user, the goal is to find other similar users whose ratings strongly correlate with the current user, and to recommend items rated highly by these similar users, but not rated by the current user. Almost all existing commercial recommenders use this approach (e.g., Amazon).

Coming back to our model of the utility matrix U, where users are represented by rows and items by columns, we can say that users are similar if their (rows or) vectors are "close" according to some distance measure (which again could be the cosine measure), and recommendation for user X is made by looking at those users that are most similar to X and then recommend items that these users like.

Obviously, there is a "dual" version of CF recommenders: Items are also similar if their representing (columns or) vectors are close. So if a rating for an item is missing, a recommender could estimate it based on rating for similar or close items, and this can use the same similarity metrics and prediction functions as before.

Thus, CF recommenders distinguish user-to-user recommendation, made by finding users with similar taste or profile, from item-to-item recommendation, made by finding items that have similar appeal to many users. Going back to the sample utility matrix U seen earlier, which has been populated further, we could conclude the following:

	MI1	MI2	MI3	JB	FF5	FF6	FF7
A	4	5		5	1	5	
B	5	5	4	2		4	
C	5			2	4	5	
D		5				5	3

Users B and C both liked movie MI1 as well as FF6 and disliked movie JB, so they might have similar tastes; thus MI2 could be a good recommendation for C. Conversely, users A and D liked both MI2 and FF6, so we may conclude that people who like FF6 will also like MI2, and hence MI2 will be recommended to user C.

Let us come back to the Jaccard similarity introduced above and consider whether it is appropriate. To determine the similarity of users B and C, we consider their associated vectors and ignore the missing entries. Thus, when B is considered as a set, we get B = {2, 4, 4, 5, 5} (technically, we need to consider multisets here, where duplicate entries are allowed, due to the fact that each number represents a distinct valuation of some item). Similarly, C = {2, 4, 5, 5}. Hence we obtain:

$$\mathrm{B} \cap \mathrm{C} = \{2, 4, 5, 5\}$$
$$\mathrm{B} \cup \mathrm{C} = \{2, 4, 4, 5, 5\}$$

which implies

$$\mathrm{sim_J}(\mathrm{B}, \mathrm{C}) = 4/5 = 0.8$$

By the same type of calculation, we obtain the following, for example, for C and D:

$$\mathrm{C} \cap \mathrm{D} = \{5, 5\}$$
$$\mathrm{C} \cup \mathrm{D} = \{2, 3, 4, 5, 5\}$$

which implies

$$\mathrm{sim_J}(\mathrm{C}, \mathrm{D}) = 2/5 = 0.4$$

While the second result (regarding C and D) is somewhat more intuitive than the first (since C and D hardly have anything comparable), we can see that the Jaccard measure is *not* appropriate in this case, since the information to which item a rating value belongs is completely lost, and hence we are somehow comparing apples and oranges. This would be even more striking if we had other users, say, E and F such that, for example, E rated MI2 with 1 and FF7 with 5, while F rates exactly opposite (and both so far rated nothing else): The Jaccard similarity of E and F would be 1, i.e., these users would be identified as having the same taste, but that would be far from valid.

A more appropriate measure in cases like these is the cosine measure described next. Since we do have a utility matrix, we can look at the various vectors contained in it. In particular, we now consider user preferences (i.e., rows in the utility matrix) as vectors in a multi-dimensional space and will look at their pairwise *cosine distance*, i.e., the angle between them. Since our vectors have positive integer components only, we are technically looking at the discrete version of a Euclidian space. Our intuition is that the smaller the angle between two vectors, the more the

point in the same direction and hence the more similar they are: Two vectors x and y with the same orientation have a cosine similarity of 1; if x and y are at 90°, they have a cosine similarity of 0.

A quick recap from any geometry book will reveal that the cosine of two vectors x and x, denoted cos(x, y) is defined as follows:

$$\cos(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

In words, the cosine of vectors x and y is calculated as the dot product of x and y divided by the L_2 norms of x and y, i.e., their Euclidean distances from the origin. The dot product of vectors $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_n]$ is defined as

$$x \cdot y = \sum_{i=1}^{n} x_i \cdot y_i$$

and the L_2 norm of vector x is

$$||x|| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

(correspondingly for vector y).

We now apply this measure sim(x, y) = cos(x, y) to our original utility matrix:

	MI1	MI2	MI3	JB	FF5	FF6	FF7
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Missing entries will be no longer be ignored (as for the Jaccard measure), but now be treated as 0, and we first look at the cosine of the angle between users A and C:

$$\frac{5 \times 2 + 1 \times 4}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{2^2 + 4^2 + 5^2}} = 0.322$$

Thus, $\text{sim}_{\cos}(A, C) = 0.322$. Similarly, we calculate cos(A, B) as

$$\frac{4 \times 5}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{5^2 + 5^2 + 4^2}} = 0.380$$

Since a larger (positive) cosine implies a smaller angle and therefore a smaller distance, this measure tells us that A is slightly closer to B than to C, confirming our intuition. By comparison, it is easily verified that, using Jaccard similarity, we would have obtained $\text{sim}_J(A, C) = 0.5$ and $\text{sim}_J(A, B) = 0.2$.

So far we have focused on user-to-user collaborative filtering, yet there is the obvious alternate ("dual") view of item-to-item collaborative filtering:

- For item *i*, find other similar items;
- Estimate a rating for item *j* based on ratings for similar items.

Clearly, this kind of filtering can use the same similarity metrics and prediction functions as in the user-to-user model.

Apparently, collaborative filtering is a powerful and efficient method, which works for any item and can deliver very relevant recommendations. The bigger the underlying database is and the more the past behaviors are recorded (i.e., the larger the utility matrix U *and* the more non-blank entries in U), the better the recommendations it can produce.

On the downside, collaborative filtering might be expensive to implement and resource as well as time-consuming. A new item that has never been purchased cannot be recommended, and a new customer who has never bought anything cannot be compared to other customers and hence not be recommended any items.

In order to cope with the various drawbacks of both content-based and collaborative filtering, the approaches can be combined into "hybrid" recommenders, which use a content-based approach to score some unrated items and then use collaborative filtering for recommendations.

A final remark in this context concerns the *size* of the problems we are looking at here. Given that sites like Amazon or Netflix have millions of users or customers, and also may have thousands or even millions of products that could be subject to recommendation, a typical utility matrix will obviously have a huge dimension. In order to be able to produce recommendations in a timely fashion, ideally while a user is navigating Amazon's or Netflix' pages, two directions of technical support need to be explored: Relevant computations might be done using parallelization and paradigms like map-reduce as discussed in Chap. 2. Alternatively, new algorithmic paradigms need to be employed, which go beyond the scope of this book. Finally a combination of the two is often the method of choice, since otherwise no timely recommendations would be possible.

3.7 Electronic Government

Developments in information and communication technologies (ICT) have been an enabler of enhanced, citizen focused services by governments around the world. *Electronic government*, or e-government, is widely regarded as a disruptor of traditional government service provision through greater citizen access, enhanced

democracy, improved information quality, and a range of governmental efficiencies. A definition of e-government that succinctly captures its scope of technology use is, "…a government's uses of ICT; particularly Web-based Internet applications, to enhance the access and delivery of government information and service to stakeholders such as citizens, business partners, public sector employees, and other governments, agencies and entities" (Shan et al. 2011).

It is important to note that e-government is much more than the provision of government-to-citizen (G2C) services; it is also concerned with government-to-government (G2G) interactions and government-to-business (G2B) transactions. Each of these will be discussed next.

Government-to-citizen (G2C) is the e-government category that includes all the interactions between a government and its citizens. For example citizens can find all the information they need on the Web; ask questions and receive answers, pay tax and bills, receive payments and documents. Electronic benefits transfer (EBT) is a well-document G2C example. It is expected that G2C e-government will benefit citizens in four ways:

- It will be easier for people to have their say.
- People will receive more integrated services because different government organizations will be able to communicate more effectively with each other.
- People will get better services from government organizations.
- People will be better informed because they can get up-to-date and comprehensive information about government laws, regulations, policies and services.

A commonly discussed application of G2C e-government is *e-democracy*. This is the use of electronic communications technologies such as the Internet in enhancing democratic processes within a democratic republic or a representative democracy. It is a political development still in its infancy, as well as the subject of much debate and activity within government, civic-oriented groups and societies around the world. E-democracy should seek to improve the democratic outcomes of the policy process, engage citizens in meeting public challenges, increase involvement in terms of numbers of participating citizens and, improve the quality and effectiveness of the democratic process.

There are important social issues associated with the adoption of G2C e-government. These include access to those who cannot afford Internet access, are elderly or perhaps disabled. There are also issues, like e-commerce in general, associated with citizen privacy and security. While these cannot be completely eliminated, as time progresses technical infrastructure will improve and citizens will become increasingly accepting of e-government as the new "normal."

Government-to-business (G2B) e-government is the category that includes interactions between governments and businesses (government selling to businesses and providing them with services, and also businesses selling products and services to government). Many of the G2C services are also relevant to businesses. These include paying taxes, receiving information and completing various types of online forms. What is unique to this category is government e-procurement. Government

is a large consumer of technology, vehicles, offices supplies etc. Using the collective size and purchasing power of government departments/agencies (often working together as a single purchasing entity) greater value for public money can be achieved than ever before.

The final category of e-government is government-to-government (G2G). This includes activities within government units and also those between governments. In this regard information sharing and process efficiency are the key benefits of G2G e-government. In terms of information sharing, there are obvious situations where timely and accurate data staring could lead to much better governmental outcomes. Such sharing might be seen between government departments such as Immigration and Inland Revenue, along with Police and Justice.

Broadly, the vision of e-government is based around enhancing public participation and by providing a progressive and reformist approach to bureaucracies (Cumbie and Kar 2016). While what has transpired in a practical sense may not fully aligned with what was initially envisioned, there was a clear expectation that e-government was not about automating existing processes, but in offering improved service delivery, integrated services, and market development (Grant and Chau 2006).

Benchmarking studies tend to categorize e-government initiatives as being at one of several distinct stages, or levels, of sophistication. The traditional view through the late-1990s was that e-government developments would parallel those being observed in the commercial world; essentially offering basic online information; then citizen-requested information; followed by extra online service channels. It was also expected that a step change in coverage might subsequently occur via extensive online collaborations with a wide range of stakeholders on the way to offering a full e-government service (after De Kare-Silver 1998). While even recent models of e-government sophistication offer evidence of such progression (e.g., Norris and Reddick 2012), the advent of Web 2.0 technologies also requires that e-government be viewed against the backdrop of (commercial) online social networking applications and services; in particular because a clear trend has emerged of users expecting to contribute and shape Web content themselves (Wirtz and Nitzsche 2013; Deakins et al. 2008). Recent research has started to uncover significant use of various social media applications in local government (e.g., Oliveira and Welch 2013) and these implementations provide transferable Web 2.0 migration paths for other government organizations to consider.

One model for e-government implementation that has stood the test of time, and remains relevant even today, is Deloitte's (2001) six-stage model. Taking a citizen-centric approach and seeking to establish long-term relationship with citizens, Deloitte's six stages are as follows:

- Stage One: Information publishing/dissemination. This early stage is characterized by individual government departments setting up their own websites.
- Stage Two: "Official" two-way transactions with one department at a time—with secure websites, customers are able to submit personal information and conduct transactions online.

- Stage Three: Multipurpose portals—customer-generic governments make breakthroughs in service delivery.
- Stage Four: Portal personalization—customers can access a variety of services at a single website.
- Stage Five: Clustering of common services—through the removal of duplication, this is where the real transformation of the government's structure starts to take place.
- Stage Six: Full integration and enterprise transformation—offers a full service center, personalized to each customer's needs and preferences.

It is without question that governments around the world are facing unprecedented opportunities and challenges regarding management of their information and interactions. While the commoditization of information and communication technologies coupled with Web 2.0 trends and technologies presents a plethora of possible solutions, the pace of change is such that few governments are able to keep up with it. Real-world e-government implementation has largely failed to match the hype associated with early predictions of government transformation. These predictions are still sound, it is just that the rate of change in the public sector is lagging that of for-profit industries.

3.8 Further Reading

Laudon and Traver (2015) is a comprehensive introduction in the area of electronic commerce. Data mining is the topic of Han et al. (2012) or Witten et al. (2016). Payne (2005) or Kostojohn et al. (2011) introduce the topic of customer relationship management.

Healthcare and DNA research, microbiology, but also other natural sciences like physics and chemistry have also always been among major producers of big data, although, as Reed and Dongarra (2015) explain, these scientific areas need to keep their eyes on both data and high-performance ("exascale") computing. MacManus (2015) is a comprehensive exposition of health tracking devices and applications. As we mentioned in Chap. 2 already, Big Data analytics is often based on statistical technique, like those discussed by Ramachandran and Tsokos (2015), Shasha and Wilson (2010), Kelleher et al. (2015), Provost and Fawcett (2013), or Alpaydin (2016).

Social network analysis is the topic of Barabasi (2016), Borgatti et al. (2013), Fouss et al. (2016), or Scott (2013). A wide field nowadays is that of opinion mining from social network data or sentiment analysis; an introduction is Liu (2015).

Online algorithms as used in online advertising go back to the work of Karp (1992); our exposition in Sect. 3.5 follows Leskovec et al. (2014), who discuss the topic in more detail. For details on search advertising, see also Leskovec et al.

(2014) or the original sources Kalyanasundaram and Pruhs (2000) as well as Mehta et al. (2005).

An introduction to recommender systems is given by Aggarwal (2016) or Agarwal and Chen (2016); the topic is also covered by Leskovec et al. (2014). We also mention the recent research papers by Lu et al. (2015) and Jannach et al. (2016).

http://www.springer.com/978-3-319-60160-1

The Web at Graduation and Beyond
Business Impacts and Developments
Vossen, G.; Schönthaler, F.; Dillon, S.
2017, XIV, 292 p. 78 illus., 64 illus. in color., Hardcover
ISBN: 978-3-319-60160-1