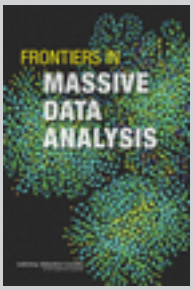

Frontiers in Massive Data Analysis

ISBN
978-0-309-28778-4

190 pages
6 x 9
PAPERBACK (2013)

Committee on the Analysis of Massive Data; Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences and Their Applications; Division on Engineering and Physical Sciences; National Research Council

2

Massive Data in Science, Technology, Commerce, National Defense, Telecommunications, and Other Endeavors

WHERE ARE MASSIVE DATA APPEARING?

Experiments, observations, and numerical simulations in many areas of science nowadays generate terabytes of data and, in some cases, are on the verge of generating many petabytes. This rapid growth heralds an era of "data-centric science," which requires new paradigms addressing how data are captured, processed, discovered, exchanged, distributed, and analyzed. While traditional methods of analysis have largely focused on analysts being able to develop and analyze data within the confines of their own computing environment, the growth of big data is changing that paradigm for many disciplines, especially in cases in which massive amounts of data are distributed across locations. The distributed and heterogeneous nature of the data provides substantial challenges for many disciplines in the physical and life sciences and also in commerce, medicine, defense, finance, telecommunications, and other industries.

The fact that scientific data sets across a wide range of fields are multiplying is an important driver for modern science. Analyses of the information contained in these data sets have already led to major breakthroughs in fields ranging from genomics to astronomy and high-energy physics, encompassing every scale of the physical world. Yet much more remains, and the great increase in scale of the data creates complex challenges for traditional analysis techniques.

It is not only experimental measurements that are growing at a rapid pace. As stated in Szalay (2011, p. 34): "The volume of data produced by computer simulations (used in virtually all scientific and engineering disci-

plines today) is also increasing at an even faster rate. Intermediate simulation steps must often be preserved for future reuse because they represent substantial computational investments. The sheer volume of these data sets is only one of the challenges that scientists must confront." Data analyses in some other disciplines (e.g., environmental sciences, wet laboratories in life sciences) are challenged to work for thousands of distinct, complex data sets with incompatible formats and inconsistent metadata.

While the scientific community and the defense industry have long been leaders in generating large data sets, the emergence of e-commerce and massive search engines has led other sectors to confront the challenges of massive data. For example, Google, Yahoo!, Microsoft, and other Internet-based companies have data that are measured in exabytes (10^{18} bytes). The availability and accessibility of these massive data sets is transforming society and the way we think about information storage and retrieval.

Social media (e.g., Facebook, YouTube, Twitter) have exploded beyond anyone's wildest imagination, and today some of these companies have hundreds of millions of users. Social-media-generated texts, images, photos, and videos comprise an unexpected and rapidly growing corpus of data. Data mining of these massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity, and national intelligence. New algorithms that assess these data in ways other than counting hits on key words, such as the analysis of social relationships, involves large graph analyses and requires new scalable algorithms.

Understanding and characterizing typical Web behavior dynamically (because the time scale of changes on the Internet is in minutes) presents remarkable challenges. In this cyber-oriented world, behavior that does not fit the patterns is often related to malware or denial-of-service attacks. Recognizing these in time, estimating the impact on human behavior, and responding is a new and emerging challenge that has few parallels in science.

Capturing and indexing the Internet has created whole sets of new industries. Some of the world's largest companies are trading in information and have built their business model on appropriately customized advertisements. Interpreting user behavior and providing just-in-time advertisements customized to the users' profiles require very sophisticated data management capabilities and efficient algorithms. Service-sector companies specializing in Internet-based auctions, like eBay or Amazon, have developed sophisticated analytics capabilities. Almost all Web-based companies today are capturing user actions, even if they do not immediately analyze them. This confluence of technologies has created a whole new industry, one based inherently on massive data.

CHALLENGES TO THE ANALYSIS OF MASSIVE DATA

A number of challenges exist in both data management and data analysis that require new approaches to support the "big data" era. These challenges span generation of the data, preparation for analysis, and policy-related challenges in its sharing and use. Initiatives in research and development that are leading to improved capabilities include the following:

- Dealing with highly distributed data sources,
- Tracking data provenance, from data generation through data preparation,
- Validating data,
- Coping with sampling biases and heterogeneity,
- Working with different data formats and structures,
- Developing algorithms that exploit parallel and distributed architectures,
- Ensuring data integrity,
- Ensuring data security,
- Enabling data discovery and integration,
- Enabling data sharing,
- Developing methods for visualizing massive data, and
- Developing scalable and incremental algorithms.

As data volumes increase, the ability to perform analysis on the data is constrained by the increasingly distributed nature of modern data sets. Highly distributed data sources present challenges due to diverse natures of the technical infrastructures, creating challenges in data access, integration, and sharing. The distributed nature also creates additional challenges due to the limitations in moving massive data through channels with limited bandwidth. In addition, data produced by different sources are often defined using different representation methods and structural specifications. Bringing such data together becomes a challenge because the data are not properly prepared for data integration and fusion, and the technical infrastructures lack the appropriate information infrastructure services to support analysis of the data if it remains distributed. Statistical inference procedures often require some form of aggregation that can be expensive in distributed architectures, and a major challenge involves finding cheaper approximations for such procedures. Finally, security and policy issues also limit the ability to share data. Yet, the ever-increasing generation of data from medicine, physical science, defense, and other industries require that analysis be performed on data that are captured and managed across distributed databases.

In addition to challenges posed by the distributed nature of most massive data, the increase of data can also limit, in other ways, the amount

of analysis that can be performed. For example, some data require high-performance computational infrastructures for data preparation before analysis can even begin, and access to such capabilities may be limited. An example would be an Earth science investigation that requires first converting the data to a common spatial grid. In cases where the analysis depends on such an expensive pre-processing step, the usefulness of the massive data is enhanced if the data-collection system is engineered to collect data of high quality in forms that are ready for analysis without the pre-processing.

However, rather than having data pre-processed for all scenarios, and thus taking up substantial storage, ad hoc investigations may require that data be processed, and thus sufficient computing infrastructures must be in place to support such ad hoc analysis. This can be desirable scientifically, because understanding of what data are needed may become clearer as an investigation proceeds. In order to support this evolutionary cycle, software systems that handle massive data must be inherently information-driven. That means that their content should be based on explicit information models that capture rich semantics that improve the provenance and understanding of the data. This in turn makes it easier to correlate distributed data sets, thus improving the ability to effectively search large collections of data without requiring changes and updates to the software (and hardware) as the data model evolves and changes during the scientific process.

Finally, challenges exist in better visualizing massive data sets. While there have been advances in visualizing data through various approaches, most notably geographic information system-based capabilities, better methods are required to analyze massive data, particularly data sets that are heterogeneous in nature and may exhibit critical differences in information that are difficult to summarize. This topic is discussed in Chapter 9.

TRENDS IN MASSIVE DATA ANALYSIS[1]

While improvements in computer hardware have enabled today's explosion in data, the performance of different architectural components increases at different rates. Central processing unit (CPU) performance has been doubling every 18 months, following Moore's Law. The capacity of disk drives is doubling at a similar rate, somewhat slower than the original Kryder's Law prediction (Walter, 2005), driven by higher density platters. On the other hand, the disks' rotational speed has changed little over the past 10 years. The result of this divergence is that while sequential input/output (I/O) speeds slowly increase with density, random I/O speeds have changed only moderately. Because of the increasing difference between the sequential and random I/O speeds of disks, only sequential disk access is

[1] The first four paragraphs of this section follow Szalay (2011).

possible—if a 100-terabyte (TB) computational problem requires mostly random access patterns, it cannot be done. Finally, network speeds, even in the data center, are unable to keep up with the increases in the amount of data. Said differently, with petabytes (PB) of data, we cannot move the data to where the computing is; instead, we must bring the computing to the data. More discussion of hardware and software for managing massive data is found in Chapter 3.

The typical analysis pipeline of a data-intensive scientific problem starts with a low-level data access pattern during which outliers are filtered out, aggregates are collected, or a subset of the data is selected based on custom criteria. The more CPU-intensive parts of the analysis happen during subsequent passes. Such analyses are currently often implemented in research environments in small clusters of linked commodity computers (e.g., a "Beowulf cluster") that combine compute-intensive, but storage- and I/O-poor, servers with network-attached storage. These clusters can handle problems of a few tens of terabytes, but they do not scale above 100 TB because they are constrained by the very high costs of petabyte-scale enterprise storage systems. Furthermore, as these traditional systems grow to meet modern data analysis needs, we are hitting a point where the power and space requirements for these systems exceed what is available to individual investigators and small research groups.

Existing supercomputers are not well suited for data-intensive computations either, because while they maximize CPU cycles, they lack I/O bandwidth to the mass storage layer. Moreover, most supercomputers lack disk space adequate to store petabyte-size data sets over the multi-month periods that are required for a detailed exploratory analysis. Finally, commercial cloud computing platforms are not the answer either, at least not today. The data movement and access fees are excessive compared to purchasing physical disks, the I/O performance they offer is substantially lower (e.g., 20 megabytes per second), and the amount of provided disk space (often in the range of, say, 10-50 gigabytes) is woefully inadequate for massive data.

Based on these observations, it appears that there is an unmet need today for capabilities to enable data-intensive scientific computations: an inexpensive yet efficient product for data-intensive computing in academic environments that is based on commodity components. The current situation is not scalable and not maintainable in the long run. This situation is analogous to the one that led to the development of the Beowulf cluster.

As data sets are growing at, or faster than, Moore's Law, they are growing at least as fast as computing power increases. This trend tends to limit analytical techniques to those that scale, at most, linearly with the number of data points (N), although those that scale as $N \log N$ are also acceptable because the $\log N$ factor can be made up through parallelism. It becomes increasingly difficult to tackle computationally challenging data analyses

using existing algorithmic tools, and there is a need to develop new tools that target near-linear computational complexity.

The large-data analytics and e-commerce companies have spent substantial resources on creating a hardware/software framework with performance that scales well with massive data. The typical approach is to create large data centers consisting of hundreds of thousands of low-end computers managed with an extreme economy of scale. Their main features include the following:

- Centralization of large-scale infrastructure,
- Data sizes at the petabyte to exabyte scale,
- Architectures that are highly fault-tolerant, and
- Computing that is collocated with the data for large-scale analytics.

While large businesses in the past have used relational databases, these do not scale well to such extreme sizes. Industries dealing with big data are reacting to data that is more distributed, heterogeneous, and generated from a variety of sources. This is leading to new approaches for data analysis and the demand for new computing approaches. Various innovative data management solutions have emerged, many of which are discussed in Chapter 3. These models work well in the commercial setting, where enormous resources are spent on harvesting and collecting the data through actions such as Internet crawling, aerial photos for geospatial information systems, or collecting user data in search engines. Some of the technical trends that have been occurring to address the data challenges include the following:

- Distributed systems (access, federation, linking, etc.),
- Technologies (MapReduce algorithms, cloud computing, Workflow, etc.),
- Scalable infrastructures for data- and compute-intensive applications,
- Service-oriented architectures,
- Ontologies, models for information representation,
- Scalable database systems with different underlying models (relation to triple stores),
- Federated data security mechanisms, and
- Technologies for moving large data sets.

Many of these technologies are being used to drive toward more systematic approaches.

As discussed in the examples below, many groups are setting up tools that support pipeline or workflow approaches to data analysis. Rather than constructing one large database, the general concept is to enable analysis by bringing together a variety of tools that allow for capture, preparation,

management, access, and distribution of data. This collection of tools is configured as a series of steps that constitute a complex workflow for generating and distributing data sets. Such a pipeline may also extract data from operational databases and systems and put that data into environments where it can be prepared and fused with other data sets and staged into systems that support analysis. Challenges include co-utilization of services, workflow discovery, workflow sharing, and maintaining information on metadata, information pedigree, and information assurances as data moves through the workflow.

Several multi-organizational groups are addressing the big-data issue and helping to bring the required scientific expertise and attention to this important problem. Examples are the groups that organize the Extremely Large Databases workshop series,[2] those working in the database research community, and those that have produced open-source technologies such as Hadoop (see Chapter 3). As the state of research in data-intensive systems is focusing on new ways to process data, these groups are leveraging open-source technologies and similar approaches to handle the orchestration of data-processing algorithms and the management of massive amounts of heterogeneous data. A major challenge in the big-data area is in evaluating ways in which distributed data can be analyzed and in which scientific discoveries can be made. Many of the aforementioned technologies, such as Hadoop, and also the proposed SciDB database infrastructure for science data management, have been co-developed across several organizations (both research laboratories and industry) and deployed across organizational boundaries to support analysis of massive data sets. The challenge is both one of developing appropriate systems as well as creating novel methods to support analysis, particularly across highly distributed environments.

In many ways, the big-data problem is simplified when data are centrally stored. However, modern data sets often remain distributed because of technical, social, political, and economic reasons, increasing the challenge to efficiently analyze the data and driving the need to build virtual data systems that integrate assets from multiple organizations. In this emerging paradigm, these virtual systems require elasticity, separation of concerns, scalability, distribution, and information-driven approaches.

Cloud computing is offering an attractive means to acquire computational and data services on an "as needed" basis, which addresses the need for elasticity. This option requires, however, that systems are architected to take advantage of such an infrastructure, which is not often the case. Many systems must be rethought from first principles in order to better exploit the possibilities of cloud computing. For example, systems that decouple data storage, data management, and data processing are more likely to rap-

[2] See the Extremely Large Databases website at http://www.xldb.org.

idly take advantage of elastic technology paradigms like cloud computing because they are more suited to leveraging the generic services that cloud computing can readily provide.

EXAMPLES

Earth and Planetary Observations

In the physical sciences domain, there is an increasing demand for improving the throughput of data generation, for providing access to the data, and for moving systems toward greater distribution, particularly across organizational boundaries. Earth, planetary, and astrophysics missions, for example, have all seen an order-of-magnitude increase in data over the past decade, rising from hundreds of gigabytes in some cases to well into the tens to hundreds of terabytes range. The Square Kilometer Array project, as an example, is predicted to produce hundreds of terabytes a second![3] Studies released by each of these disciplines in their decadal reports suggest that this trend will continue. Many Earth and planetary missions have instruments that will continue to generate observations of massive size that will substantially increase the scientific archives worldwide. Climate research also continues to grow at a rapid pace as climate models and satellite observations grow and are needed to support new discovery. NASA's Earth Science enterprise, for example, now manages data collections in the several-petabyte range. In planetary science, the amount of data returned from robotic exploration of the solar system between 2002 and 2012 was 100 times the amount of data collected from the previous four decades. Astrophysics has seen similar increases, and all of these disciplines have experienced a continued increase in the geographic distribution of the data.

As already noted, the increase of data within these science environments has, in many cases, led to the construction of data "pipelines" that acquire data from instruments, process and prepare that data for scientific use, capture the data into well-engineered data management systems, and then provide ad hoc services for data distribution and analysis. Such infrastructures have required advanced computing capabilities that often span multiple institutions and support well-orchestrated information services.

While, traditionally, many of these workflows are constructed for each instrument or mission, there is an increasing interest in performing analysis across data sets that may span different instruments or missions, even from

[3] This unprecedented increase will require innovation beyond our current understanding of computation and storage over the next 10 years to achieve the project's ambitious requirements and science goals. Background on the Square Kilometer Array project may be found at http://www.skatelescope.org.

different disciplines. This type of data integration and inter-comparison is not limited to just observational data sets. Within the climate research community, effort is under way to prepare observational data so that it can be inter-compared against the output of climate models. However, within the climate discipline, as with others, data from different institutions, systems, and structures use different standards and measurements, and this lack of standardization makes such analysis difficult, due to the heterogeneity.

Astronomy[4]

Astronomy is a good example for studying how the data explosion happened and how long it might continue. In this realm, successive generations of exponentially more-capable sensors, at the same cost, are the reason for the data explosion (all being traceable to advances in semiconductor technology and, ultimately, to Moore's Law). New generations of inexpensive digital cameras come out every 6 months, and satellites have ever-higher resolution and more pixels. Even old telescopes are getting new instruments that collect more data. For example, the Dark Energy Survey is building a huge array of charge-coupled devices (CCDs) to be placed on an older telescope in Chile.

However, not every domain of science has such growth areas. One could argue that optical astronomy will soon reach the point when increasing the size of collector arrays will become impractical, and the atmospheric resolution will constrain the reasonable pixel size, causing a slow-down in data collection. But time-domain astronomy is emerging, and by taking images every 15 seconds, even a single telescope (e.g., the Large Synoptic Survey Telescope) can easily generate data that can reach 100 PB in a decade.

New instruments and new communities emerge to add to the big-data movement. Radio astronomy, with focal plane arrays on the horizon, is likely to undergo a paradigm shift in data collection, resembling the time when CCDs replaced photographic plates in optical telescopes. Amateur astronomers already have quite large, cooled CCD cameras. When a community of 100,000 people starts collecting high-resolution images, the aggregate data from amateur astronomers may easily outgrow the professional astronomy community.

Biological and Medical Research

A substantial amount of analysis is being performed using data collected by medical information systems, most notably patient electronic health records. This information represents a wealth of data both to im-

[4] This section is adapted from Szalay (2011).

prove individual health-care decisions as well as to improve the overall health-care delivery enterprise. The U.S. Food and Drug Administration, for example, is building an active drug safety surveillance system utilizing data from de-identified medical record databases. Medical researchers are gathering together to share information about interventions and outcomes in order to perform retrospective analysis, and insurance companies continue to mine data to improve their own models.

The genomics revolution is proceeding apace, with the cost of sequencing a single human genome soon to drop below $1,000. As the cost decreases, it becomes feasible to sequence multiple genomes per individual, as is being envisaged in tumor genomics initiatives. Overall, data volumes in genomics are growing rapidly. The Short Read Archive at the National Center for Biological Information is soon expected to exceed a petabyte. As more and more high-throughput sequencers are deployed, not just in research but also in hospitals and other medical facilities, we will see an even faster data growth in genomic information.

Neuroscience is increasingly using functional magnetic resonance imaging, where each session can easily result in tens of terabytes of data. Longitudinal studies of hundreds of patients generate data measured in petabytes today. Studies of the cardiovascular systems are creating multiscale simulations from the molecular scale to those of the human circulation system. The European Human Brain Project is setting out to integrate everything we know about the brain in massive databases and detailed computer simulations. And ultra-high-resolution microscopy is also generating very large data sets in cell biology.

Large Numerical Simulations[5]

Numerical simulations are becoming another new way of generating enormous amounts of data. This has not always been the case. Traditionally, these large simulations (such as gravitational *N*-body simulations in astrophysics or large simulations of computational fluid dynamics) have been analyzed while the simulation was running, because checkpointing and saving the snapshots was overly expensive. This fact traditionally limited the widespread use of simulation data.

Even when a few snapshots have been saved and made public, downloading large files over slow network connections made the analysis highly impractical once simulations reached the terabyte range. The Millennium simulation in astrophysics has changed this paradigm by creating a remotely accessible database with a collaborative environment, following the example of the Sloan Digital Sky Survey SkyServer. The Millennium data-

[5] This section is adapted from Szalay (2011).

base drew many hundreds, if not thousands, of astronomers into analyzing simulations as easily as if the observational data were publicly available.

The emerging challenge in this area is scalability. The Millennium has 10 billion particles. The raw data is about 30 TB, but the 1 TB database does not contain the individual dark matter particles, only the halos, subhalos, and the derived galaxies. Newer simulations are soon going to have a trillion particles, where every snapshot is tens of terabytes, so the data problem becomes much worse. At the same time, there is an increasing demand by the public to get access to the best and largest simulations; it is inevitable that the Millennium model is going to proliferate. There may be a need for a virtual observatory of simulations that can provide adequate access and the ability to do analysis, visualization, and computations of these large simulations remotely. This need cuts across all disciplines, because simulations are becoming more commonplace in all areas of physical science, life sciences, economics, and engineering.

As data become increasingly unmovable, the only way to analyze them is "in place." Thus, new mechanisms are needed to interact with these large simulations, because it is not feasible to simply download raw files. For interactive visualizations, it would be easier to send a high-definition, three-dimensional video stream to every interested scientist in the world than it would be to move even a single snapshot of a multi-petabyte simulation from one place to another.

New and interesting paradigms for interacting with large simulations are emerging.

In a project related to isotropic turbulence, data are accessed via a Web service where users can submit a set of about 10,000 particle positions and times and then retrieve the interpolated values of the velocity field at those positions. This can be considered as the equivalent of placing small "virtual sensors" into the simulation instead of downloading all the data or significant subsets of it. The service is public and is typically delivering about 100 million particles per day worldwide. Several papers appearing in the top journals have used this facility.

Telecommunications and Networking

Managing a modern globe-spanning highly reliable communications network requires extensive real-time network monitoring and analysis capabilities. Network monitoring and analysis is used for tasks such as securing the network from intruders and malefactors, rapid-response troubleshooting to network events, and trend prediction for network optimization.

Large telecommunication providers such as AT&T and Verizon offer a wide range of communication services, ranging from consumer (mobile voice and data, wired broadband, television over Internet protocol, and

plain old telephone service) to business (data centers, virtual private networks, multiprotocol label switching, content caching, media broadcast), to a global long-haul communications backbone. Each of these services contains many interacting components, and understanding network issues generally requires that data about the interacting components be correlated and analyzed. Further, the global network contains tens of thousands of network elements distributed worldwide.

In addition to the large volumes of data involved, the major problem in telecommunications and networking data analysis is the complexity of the data sets (hundreds to thousands of distinct data feeds) and the requirement for real-time response. Monitoring the health of tens of thousands of backbone routers generates large data sets, but backbone routers are only one component of the global end-to-end communications network. Many user applications—for example, streaming music to a smartphone—require the correct and efficient operation of dozens of different network elements. Troubleshooting a bad connection requires that dozens of data feeds be monitored and correlated.

Responding to network events, such as misbehaving routers, stalled routing convergence algorithms, and so on, or to network intruders, requires fast response; delays result in customer dissatisfaction, or worse. Therefore, all of the hundreds of data feeds must be managed in a real-time warehouse, which can provide timely answers to troubleshooting queries. The technology for such real-time warehousing is still being developed.

Social Network Analysis

Social network analysis is the science of understanding, measuring, and predicting behavior from a relational or structural perspective. Using a blend of graph-theoretic and nonparametric statistical techniques, researchers in this area take data, such as phone data or observations indicative of interpersonal connections, and identify who are the key actors and key groups within and across networks, also identifying special patterns and important pathways. Traditional data sets focused on information within groups—such as who worked with whom or who is friends with whom—or between groups, such as which countries are allied or which organizations supplied goods to which others. The network in those traditional investigations might be a simple one- or two-mode network (i.e., a network with nodes that fall into just one or two classes), and the links were often binary. Often the data were from a small contained group.

Today the field of social network analysis has exploded, and data often take the form of meta-networks in which information about who, what, where, how, why, and when are linked using multi-mode, multi-link, multi-level networks. (See, e.g., Carley, 2002.) The links are probabilistic,

and the nodes have states that may change over time. The size of networks of interest are often larger than in the past, such as the entire citation network in the web-of-science or the network of phone calls in all countries over a period of 12 months. The availability of, and interest in, such massive network data has increased as social media sites have become more prevalent; as data records for public functions (such as home sales records and criminal activity reports) have increasingly been made public; and as various corporations make such data available, at least in anonymized form (such as phone records, Internet movie databases, and the web-of-science).

Massiveness for network data arises on several fronts. First, the number of nodes in data of interest has grown from hundreds to several million or billion. Second, the number of classes of nodes that need to be included in a single analysis has grown. For example, web-of-science data have authors, topics, journals, and institutions, with each of these "nodes" having multiple attributes. Third, the data are often collected through time and/or across regions. For example, Twitter, Lexis-Nexis, AIS, and various sensor feeds all have networks embedded in space and time. In particular, social media contain meta-network data that are massive and still growing.

Each source of massiveness presents the following technical challenges:

- The increase in the number of nodes whose data are being analyzed means that many of the traditional algorithms must be replaced by ones that scale better. This has been easy for metrics that rely only on measures associated with individual nodes and their direct links to other nodes—i.e., on local information for each node. For example, degree centrality, the number of links to/from a node, scales well, and it also can be readily adapted to streaming data. However, algorithms that do not scale well are those (such as betweenness) that rely on examining paths through an entire network or on local simulations that use multi-mode, multi-link data. In general, many single-node algorithms that include key node identification and grouping algorithms either scale well, are already parallelized, or have heuristic-based approximation approaches (Pfeffer and Carley, 2012). The main challenges here are rapid re-estimation given streaming data, estimating emergence and degradation of a node's position over time given dynamic data with missing information, comparison of multiple networks, and enumeration of motifs of interest.
- Multi-mode data are challenging in that there are few metrics, and new ones are needed for each application. To be sure, a set of metrics exists for two-mode networks; however, most of the massive data is n-mode. From a massive data perspective, the key challenge is that the search paths tend to increase exponentially

with the number of node classes (i.e., modes). Improved metrics, scalable multi-mode clustering algorithms, improved sets of interpretations, and improved scaling of existing metrics are the core challenges.

- For temporal data there are two core challenges: incremental assessment and atrophication/emergence. Incremental assessment requires new algorithms to be defined that reflect social activity, which can be rapidly computed as new data become available. Even newly developed incremental algorithms are still exponential with network size; more importantly, it is not clear that the existing metrics (e.g., betweenness), even if sufficiently scalable incremental algorithms were to be developed, are meaningful in truly massive networks. Other challenges center around the problem of identifying points of atrophication and emergence, where portions of the network are fading away or emerging. The identification of simple temporal trends is not particularly a challenge, as Fourier analysis on standard network metrics provides guidance and scales well. With temporal data such as email and Twitter, not all nodes (the people who send information) are present in every time period. Understanding whether this lack of presence represents missing data due to sampling, temporary absence, or to a node actually leaving the network is a core challenge.

Thus, the harder problems that are particularly impacted by massive data include (1) identifying the leading edge in a network as it is being activated (e.g., who is starting to contract a disease or where is a revolution spreading), (2) identifying what part of the network is contributing to anomalous behavior, and (3) updating metrics as data changes. Geo-temporal network data present still further challenges, due in part to the infrastructure constraints that inhibit transmitting and sharing geo-images and the lack of large-scale, well-validated, spatial data for locations of interest in network analyses. However, even if these technical and data problems were solved, dynamic geo-enabled network analysis would still be problematic due to the lack of a theoretical foundation for understanding emergence in spatially embedded networks. A major problem in this area is diffusion. Well-validated spatial and social network models exist for the spread of distinct entities such as disease, ideas, beliefs, technologies, and goods and services. However, these models often do not make consistent predictions; that is, spatial and social network models disagree, and they typically do not operate at the same scale. Having an integrated spatial-social-network model for the diffusion of each type of entity is critical in many areas. The increase in geo-temporal tagged network data is, for the first time, making it possible to create, test, and validate such models; how-

ever, current mathematical and computational formulations of these models do not scale to the size of the current data.

Much current network analysis is done on individual standalone single-processor machines. However, that is changing. There are a few tools (e.g., the Organizational Risk Analyzer (ORA) toolkit) that have parallelized the algorithms and make use of multi-processors. In addition, there are Hadoop versions for some of the basic algorithms,[6] thus enabling utilization of cloud computing. Many diffusion routines can now be run on Condor clusters. In general, there is a movement to distributed computing in this area; and although the trend will continue, the existing technologies for network analysis in this area are in their infancy. New tools are appearing on a regular basis, including special tools for supercomputers, tools that take advantage of special processors, chips with built-in network calculations, and algorithms that utilize the memory and processors in the graphics display.

Social networking—the use of a social media site to build, maintain, review, and disseminate information through connections—is a growing trend. Such sites include Twitter, Facebook, LinkedIn, and YouTube, and they are a growing source of massive data. In general, social media are being applied as a means to gather, generate, and communicate data in a rapid fashion. They are used in marketing by companies to announce products and collect consumer feedback and also by companies to discern other companies' secrets. They are used by groups to organize social movements, protests, track illegal activity, record the need for social services (e.g., pothole filling and snow removal), provide feedback on quality of restaurants, hotels, services, and safety of areas, and so on. Most technologies applied to such social media do little more than collect data—sometimes providing simple visualizations—and simply count the frequency of messages, key words, hashtags, etc. Truly making use of these data requires scalable clustering techniques, real-time ontology abstraction, and on-the fly thesauri creation for extracting the complete network associated with a topic of interest.

Social network analysis technologies can be used to assess social media data, while social network theory can be used to address how people will connect via social media and how it will change the nature of their interactions. However, many challenges remain. Social network analysis has traditionally focused on small, complete networks (i.e., fewer than 100 members, in which all members were contacted), where interaction can be face-to-face, and all data come from one time period. To exploit social media data, techniques have to be expanded to handle large networks (e.g., thousands or millions of nodes), where the data are sampled rather than

[6] See, for example, X-Rime, available at http://xrime.sourceforge.net.

complete (so there may be sampling biases in what data are captured), and for which the data are typically dynamic and the dynamics are of interest. Hence, issues of missing data, link inference, bias, forecasting, and dynamics are now of great interest. Also, network metrics are highly sensitive to missing and erroneous data, and so alternative metrics, confidence intervals on existing metrics, and procedures for inferring missing data are all needed. Advances in these areas are occurring rapidly. An emerging challenge, however, is how one can cross-identify the same person in multiple social media. This is particularly important for tracking criminal activity, terrorists, or pedophiles.

Social media are still evolving. As they mature, the shape of the technology itself will be different, and new users that will have grown up with them will be the dominant group. As such, cultural norms of usage will emerge. Changes in security and privacy options on social media sites are liable to make such cross-identification even more challenging than it is currently. Because the technology is a major source for collecting and processing massive data, the needs and challenges facing data analysis are likely to change as the technology matures.

Some of the common questions of analysis with respect to social networks include the following:

- *Effective marketing with social media.* How can companies and governments measure the effectiveness of their social media campaigns, assess change in the resulting culture, and understand when the message changes what people do?[7] What new scalable social network tools, techniques, and measures are needed for identifying (1) key actors for spreading messages, (2) early adopters, (3) the rate of spread, and (4) the effectiveness of the spread, given the nature of social media data, to track crises and identify covert activity?[8]
- *Sentiment assessment monitoring and control.* How can one measure, assess, forecast, and alter social sentiment using diverse social media? What new scalable social-network techniques are needed for assessing sentiment, identifying sources of sentiment, tracking changes in groups and sentiment simultaneously, determining whether the opinion leaders across groups are the same or different, and so on? (Pfeffer et al., Forthcoming).
- *Social change in images.* A vast amount of the data in social media sites is visual—videos and photographs. How can these data be

[7] See, e.g., Minelli et al. (2013).

[8] See, e.g., De Choudhury et al. (2010), Wakita and Tsurumi (2007), and Morstatter et al. (2013).

assessed, measured, and monitored in a scalable fashion so as to allow social change to be tracked and individuals identified through a fusion of verbal and visual data? (Cha et al., 2007).

- *Geo-temporal network analysis.* An illustrative problem is the assessment, by area within a city and time period, of the Twitter network during disasters in order to rapidly identify needs and capabilities. Currently there are few geo-network metrics; however, the increasing prevalence of geo-tagged data is creating the need for new metrics that scale well and support drill-down analysis on three dimensions at once—space, time, and size of group. Current clustering algorithms scale well on a single dimension, but scalable algorithms are needed that cluster in space and time. Finally, most spatial visualization techniques take the form of heat maps, which are often too crude to visually convey nuances in the data. But if the fullness of big data is exploited, then the overlay of points on a map becomes confusing, so new visual analytic techniques are needed (Joseph et al., 2012).

The sheer size and complexity of data about social networks, cultural geography, and social media are such that systems need to be designed to meet three goals: automation, ease of use, and robustness. Analysts simply do not have the time to capture and run even basic analyses in a time-sensitive fashion. Hence, the data-collection and analysis processes need to be automated so that information extraction can operate independently, and basic statistics identifying key actors and groups can be continually updated. Transparency must be maintained, and the analysts must be able to check sources for any node or link in an extracted network. An increasing number of jobs require tracking information using social network data, and an increasing number of activities that individuals engage in can be discerned from information on the individual's social network, particularly when multiple networks, multiple types of nodes, and multiple relations can be overlain on one another.

National Security

The rise of the Internet has enormously increased the volumes of data potentially relevant to counterterrorism, counter-proliferation, network security, and other problems of national security. Many problems in national defense involve flows of heterogeneous, largely unstructured data arriving too rapidly to be aggregated in their entirety for off-line analysis.

As an example, consider the problem of computer network defense, or cyberdefense. Network-based attacks on computer systems pose threats of espionage and sabotage of critical public infrastructure. If one can observe

network traffic, one may wish to know when an attack is under way, who is conducting the attack, what are the targets, and how a defense may be mounted. Both real-time and post-hoc (forensic) capabilities are of interest. The metadata associated with the traffic—e.g., entity X communicated with entity Y at time T using protocol P, etc.—can be regarded as a dynamic graph or arrival process whose analysis may be useful.

To cope with such problems, advances are needed across the board, from statistics to computer system architecture, but three areas can be highlighted. First, streaming algorithms that can process data in one pass with limited memory are clearly important. Second, for data at rest, transactional databases are generally not needed, but highly usable systems for hosting and querying massive data, including data distributed across multiple sites, will be essential. The MapReduce framework, discussed in Chapter 3 of this report, is perhaps a good first step. Third, better visualization tools are also needed to conserve the scarce and valuable time of human analysts.

Two areas in national security that are particularly impacted by massive data are the potential capability for the remote detection of weapons of mass destruction and improved methods of cyber command and control. Although many details of national security problems are classified, approaches to these problems often parallel current efforts in academia and industry.

REFERENCES

Carley, K.M. 2002. Smart agents and organizations of the future. Chapter 12, pp. 206-220 in *The Handbook of New Media* (L. Lievrouw and S. Livingstone, eds.). Sage, Thousand Oaks, Calif.

Cha, M., H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. 2007. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. Pp. 1-14 in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (IMC '07). ACM, New York, N.Y.

De Choudhury, M., Y.-R. Lin, H. Sundaram, K.S. Candan, L. Xie, and A. Kellihe. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Available at http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1521/1832.

Joseph, K., C.H. Tan, and K.M. Carley. 2012. Beyond "local," "social," and "category": Clustering Foursquare users using latent "topics." In *4th International Workshop on Location-Based Social Networks* (LBSN 2012) at UBICOM, September 8, 2012, Pittsburgh, Pa.

Minelli, M., M. Chambers, and A. Dhiraj. 2013. *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends*. John Wiley and Sons, Hoboken, N.J.

Morstatter, F., J. Pfeffer, H. Liu and K.M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of International AAAI Conference on Weblogs and Social Media* (ICWSM), July 8-10, Boston, Mass.

Pfeffer, J., and K.M. Carley. 2012. k-centralities: Local approximations of global measures based on shortest paths. Pp. 1043-1050 in *Proceedings of the WWW Conference 2012*. First International Workshop on Large Scale Network Analysis (LSNA 2012), Lyon, France. Available at http://www2012.wwwconference.org/proceedings/index.php.

Pfeffer, J., T. Zorbach, and K.M. Carley. 2013, Forthcoming. Understanding online firestorms: Negative word of mouth dynamics in social media networks. *Journal of Marketing Communications*.

Qualman, E. 2013. *Socialnomics: How Social Media Transforms the Way We Live and Do Business*. John Wiley and Sons, Hoboken, N.J.

Szalay, A.S. 2011. Extreme data-intensive scientific computing. *Computing in Science and Engineering* 13(6):34-41.

Wakita, K., and T. Tsurumi. 2007. Finding community structure in mega-scale social networks. Pp. 1275-1276 in *Proceedings of the 16th International Conference on World Wide Web* (WWW '07). ACM, New York, N.Y.

Walter, C. 2005. Kryder's Law. *Scientific American*, August. pp. 32-33.