Information and Influence Propagation in Social Networks

Chapter 6: Data and Software

Wei Chen, Laks V.S. Lakshmanan, Carlos Castillo

This is the authors' version of Chapter 6 only, before the review by the editor of the series. It does not include a number of corrections and changes done during the editorial process.

The full version appears in Synthesis Lectures on Data Management,

Morgan & Claypool Publishers, October 2013

http://www.morganclaypool.com/doi/abs/10.2200/S00527ED1V01Y201308DTM037

http://dx.doi.org/10.2200/S00527ED1V01Y201308DTM037

CHAPTER 6

Data and Software for Information/Influence Propagation Research

Research on information and influence propagations is motivated by applications in domains of practical importance. The availability of real-world datasets from such applications, or datasets that closely resemble them, is of utmost importance. Data is important to validate models about how viral phenomena unravel in practice under a variety of conditions.

This chapter describes classes of datasets that have been used by authors in this field. It also provides pointers to specific instances of such datasets. Finally, it briefly outlines packaged software that researchers have produced to be used by other researchers – while there are not many examples, as this research field progresses we expect more tools will be developed.

6.1 TYPES OF DATASETS

Datasets for this research must contain the elements that the models described in Chapter 2 require. First, a *social graph* $G = (V, E)$ representing social connections. Second, a family of ℓ action traces $S_t^{(i)}$, where $i = 1 \ldots \ell$ and $S_0^{(i)}$ is the seed set of action i, and $S_t^{(i)}$ for $t \geq 1$ is the set of active nodes for action i at time t.

Social networks. While the models we have described do not require symmetrical social networks, in practice many online social networks are symmetrical. This is the case of platforms where users must "confirm" or "acknowledge" that they are "friends" which other users who invite them to be part of their connections.

In some cases, social networks are not explicitly recorded by the platforms from which the data originates, or can not be directly observed. In those cases, it is common to infer the connections from interactions between users, e.g., whenever two users interact beyond a certain threshold, exchanging more than k messages, they are considered to be connected in G.

Action traces. Any change that can be observed in the state of individuals in the network can be considered an action. Sometimes these actions are explicit choices of individuals, such as deciding to switch to a different mobile phone provider (e.g. Dasgupta et al. [2008]).

A variety of other state changes can also be considered as actions, such as catching a disease or gaining weight (e.g. Christakis and Fowler [2007b]).

Real datasets describing propagation phenomena may or may not include *attribution* information. Attribution information is a family of functions $P_t^{(i)} : S_t^{(i)} \rightarrow \mathcal{P}\left(\cup_{t'<t} S_{t'}^{(i)}\right)$ for $t \geq 1$ such that $(u, v) \in P_t^{(i)}$ if and only if node v was among the ones that triggered the node u activated at time t. In this case v is said to be a *parent* of u for action i at time t. In most datasets that include attribution information, for any given action there is a single parent for every active node.

Currently, publicly-available datasets containing both a social network and action traces are relatively rare. Researchers typically resort to using datasets for which one of the two parts is missing and need to be inferred or synthesized.

Another option is to use proprietary or closed datasets, however, the exclusive use of proprietary datasets is being shunned upon in recent years. First-tier conferences including KDD and SIGMOD among others, are including increasingly stronger language in their call for papers, asking for reproducibility of research results. This does not prevent researchers from using proprietary datasets, but it often conditions that on the existence of alternative public datasets in which at least part of the experimental results can be reproduced.

6.2 PROPAGATION OF INFORMATION "MEMES"

The most common type of dataset used for research on information propagation corresponds to textual corpora from which propagation of information "memes" can be inferred. "Meme" is a term coined by Dawkins [1990] to indicate a cultural entity: an idea, behavior, or style, that an observer might consider a *replicator*. The specific form that observable "memes" can take in real-world datasets depends on the platform/system from which data are collected by researchers.

6.2.1 MICROBLOGGING

Microblogging is a well-established paradigm for online social networking sites, in which users are able to upload small pieces of content (links to web pages, photos, videos, or short pieces of text) that can be voted upon and/or "shared" by other users. Microblogging ocassionaly generates information cascades in which the number of people exposed to a given piece of content is large –the content is said to "go viral" in social media marketing parlance.

The social network in microblogging platforms is typically non-symmetrical, as users are encouraged to *follow* other users, who may or may not want to reciprocate this con-

nection. Examples of highly-visited microblogging platforms (according to Alexa[1]) include Facebook, Twitter, LinkedIn, Sina Weibo, and Pinterest, among many others.

The seed set $S_0^{(i)}$ for an action i corresponds to people who first posted a content item into what is known as their *timeline*. Every time a user u logs into a microblogging platform, s/he is shown a subset of items posted into the timelines of users that are being followed by u. In addition to browsing this content, there are simple (often one-click) mechanisms for *re-posting* an existing message.

In this setting, action traces originated by cascades of re-posts are often explicitly attributed. In practice, however, the linkage between a user posting an item and its parent is not present in the publicly-available data. This happens often with third-party clients for online social networking sites that tend to generate a new post instead of a re-post linked to the parent element – such as in the case of third-party clients for Twitter.

Sampling issues. The availability of APIs to access microblogging data has been one of the factors that have boosted the development of this area, but often these datasets are not complete. For instance, as of 2013, Twitter offers through its free API a uniform random sample of 1% of the tweets. This is problematic for researchers, as given this sampling method the information of any large cascade is almost certainly incomplete. There are methods to alleviate this problem, such as the one described by Sadikov et al. [2011].

6.2.2 NEWSPAPERS/BLOGS/ETC.

"Memes" can also be extracted automatically from textual corpora. Tracking their spread can be made easier if there are explicit links (citations/references) among documents. For instance, academic papers contain citations to other papers –those citations can be interpreted as explicit attributions of the propagation of an idea.

However, outside academic research, citations tend to be rare. For instance, blog postings –and other online media such as online news articles and general web pages– rarely cite the previous works they build upon, as shown by Adar and Adamic [2005]. In these cases, attribution for propagations needs to be inferred from other factors, for instance by taking into account the timestamp of the documents and assumming that a meme can propagate only from an older document to a newer document.

But even when propagations are explicitly attributed, characterizing computationally memes can be a daunting task. In the case of media, for instance, sometimes memes can be obvious, as in the case of an emerging popular music star whose name suddenly appears in many news articles in a collection. Other times, memes can be subtle, such as changes in the way a certain subject is framed; an interesting case in the USA is the evolution of "waterboarding" from "torture" to "enhanced interrogation technique" documented by Desai et al. [2010].

[1] `http://www.alexa.com/`

A straightforward solution is to define a "meme" as a small set of words or a phrase, as in Gruhl et al. [2004a]. Unfortunately, in practice this often yields a mixture of actions that are either too broad or too narrow. For instance, if we select news articles containing the phrase "French president" we may be selecting multiple stories that are unrelated to each other, as a person may be involved in a number of topics at the same time. On the other hand, if we select news articles containing "nuclear disarmement", we may miss related documents containing other terms such as "nuclear proliferation", etc.

Something similar happens in the case of social media that supports free-form tagging for content (e.g., with "hashtags") or URL links embedded in the content. With some exceptions, tags tend to be too broad, while URLs tend to be too narrow, as the same meme/story/event may be represented by more than one URL (Wu et al. [2011]). To overcome these problems, methods such as LDA (latent dirichlet allocation) can be applied to detect topics in collections of documents or social media postings, as in Zhao et al. [2011].

Alternatively, a solution that has been proposed for detecting memes is to start by locating "bursty" (Kleinberg [2002]) words or phrases, i.e., words or phrases that increase significantly in frequency at a given time. Once such an example is found, a method for improving precision can be applied, e.g. by demanding every document to include one extra word or phrase from a set of related terms, as in the system described by Mathioudakis and Koudas [2010].

Another solution is to consider that at every timestep, a meme can suffer small modifications. This method has been applied to the tracking of online petition letters (and their variants) as studied by Liben-Nowell and Kleinberg [2008]. An influential work in this space is the *meme tracker* system, described by Leskovec et al. [2009]. Meme tracker infers likely paths of transmission for a set of memes, starting from a list of timestamped documents, and assuming that memes can "mutate" every time they are copied.

6.3 PROPAGATION OF OTHER ACTIONS

Besides memes, there are a variety of other types of actions for which propagation data is sometimes available.

6.3.1 CONSUMPTION/APPRAISAL PLATFORMS

The consumption of products has always had a social component, as attested by the numerous trends and fashions in clothing and other industries. Online shopping is increasingly becoming a "social" activity, a phenomenon that also affect brick-and-mortar shops as many consumers go online to find recommendations for products they intend to buy [Hu and Liu, 2004]. Websites such as Pinterest `http://www.pinterest.com/` among many others, allow users to browse and *curate* collections of products and interests.

Many users also share publicly –and sometimes automatically– the media products they consume. This is done through sites like `http://kindle.amazon.com/` and `http://`

`goodreads.com` for books, `http://getmiso.com/` and `http://intonow.com` for TV shows, and `http://last.fm/` for music, among many others.

These online platforms currently include social networking features that allow users to "follow" a set of users, and receive notifications when one of those users listens to a new album or reads a new book. Users are thus exposed to lists of items consumed by other users, as well as ocasional ratings and reviews posted by them.

In this context, consuming or appraising a product becomes an action that can generate a propagation cascade. The links among users plus the sequence of products each user consumes, and/or posts an opinion on, constitute a rich (and potentially profitable) dataset to study propagation phenomena.

For instance, Bhagat et al. [2012a] analyze data about films from Flixtr and data about music preferences from Last.fm. Huang et al. [2012] use data from a Chinese media appraisal platform named Douban and the GoodReads book-reading social network.

In most cases, propagations are implicitly attributed –we do not know specifically who was the cause that user consumed a product, we just know the temporal sequence of these actions. However, in some cases there is evidence that allows to track attribution, such as in platforms that have an explicit "recommend to a friend" feature. This is the case of *@cosme*, a platform for recommendation of cosmetic products whose data has been used for research by Matsuo and Yamamoto [2009].

6.3.2 USER GENERATED CONTENT SHARING/VOTING

Platforms to share user generated content often allow some mechanisms for social networking. For instance, photo sharing site Flickr allows users to declare (non-symmetrically) people to be "family" or "friends". In this specific case, Cha et al. [2009] studied whether these links affect the decision of adding a photo as a "favorite". Anagnostopoulos et al. [2008a] studied whether the adoption of a certain tag by users when describing their photos affects their social connections. Something similar happens in user voting sites such as Digg or Reddit where there are explicit friendship links and the main action is to vote on content, as studied by Lerman [2007].

6.3.3 COMMUNITY MEMBERSHIP AS ACTION

There are many online platforms that allow users to form online groups or communities. Some researchers have considered the action of becoming a member of a group to be influenced by the social network connections of users. The hypothesis is that there is to some extent a "bandwagon effect", in which if many of the connections/friends of a user join a community, the user is also likely to join.

For instance, in blogging platform Livejournal, a user is a blogger who can declare that s/he wants to "join" a community around a topic. Such actions has been correlated with the actions of connections/friends in the same blogging platform Goyal et al. [2011b].

The definition of community can be quite flexible. In the academic world, a "community" may be represented by a journal, a conference, or even a topic. Publishing a paper on that community means becoming a member, and the social network can be inferred from past co-authorship relationships. This type of dataset was used by Backstrom et al. [2006], using data from DBLP, and Tan et al. [2010], using data from ArnetMiner.

A community can also be a group of readers of a website, such as the people who follow a particular blog through an RSS feed. This is the case of the research by Java et al. [2007] performed on blog reading platform *bloglines* `http://www.bloglines.com/`. The social network are explicit ties among readers, and an action is to subscribe to an RSS feed.

6.3.4 CROSS-PROVIDER DATA

Large Internet companies have user bases in the order of tens or hundreds of millions of users. When those companies offer a variety of services to users, the activity records from these users on the different services can be linked through site-wide user-ids. This allows them to extract the network from connections among users in one product, and track their actions in another product.

In Singla and Richardson [2008] researchers used Microsoft data from instant messaging platform MSN Messenger, together with searches in the Microsoft search engine Bing. Their conclusion was that users who are friends in the messaging platform do tend to issue similar queries. Something similar was done by Goyal et al. [2008] who used the social network of Yahoo! Instant Messenger users, and joined it with information about movies appraised by the users in Yahoo! Movies.

Using cross-provider data may have at least two caveats. First, even if the terms of services allow for it, some users may be surprised or turned off by the fact that their actions in two different (from their perspective) platforms are being correlated. Second, given that the same users can have multiple online *personas*, the nature of the platforms should be similar, e.g. if the social networking platform is oriented mostly to entertainment or gaming, the other platform should also have the same tone (and not be, e.g. a professional-oriented site).

6.3.5 PHONE LOGS

The mining of logs from (mobile) phone conversations has been mostly motivated by one specific application: reducing "churning", which is short-term switching among phone providers. In this setting, the social network is inferred by phone calls, i.e.: two users are connected if they exchange many phone calls, and the action is "switching to another phone provider" or ceasing to use the phone network, as studied by Dasgupta et al. [2008].

Despite strong privacy protections around this type of data, some phone datasets are available. For instance, Nokia offers a dataset for academic labs (not industrial ones)

http://research.nokia.com/page/12000. Another dataset, more publicly available, covers exchanges of phone calls and SMS messages among 5 million users in Ivory Coast http://www.d4d.orange.com/home.

6.4 NETWORK-ONLY DATASETS

In the previous sections we have included several datasets in which the action traces are available but the social network needs to be inferred. In this section we outline datasets for which the social network is available and actions traces need to be synthesized. Despite the obvious drawbacks of using synthetic action traces, the validation of some methods in the literature has required large-scale networks for which real action traces are not available.

6.4.1 CITATION NETWORKS

Citation networks can be extracted from a variety of sources, including academic publication repositories and patent repositories. Patents are particularly attractive because they include as many citations as academic papers, and have the additional benefit that in jurisdictions like the USA the documents are in the public domain – unlike most scientific articles.

The blogosphere (the Web of blogs), linked data repositories (such as Freebase and DBPedia), and the entire Web have also been used as citation networks. Large collections of this type are widely available; for instance Amazon offers currently a crawl of the Web containing 5×10^9 pages.

6.4.2 OTHER NETWORKS

Network data are widely available across a variety of domains.

In the domain of transportation, there are multiple publicly-available datasets containing roads, railways, or air travel routes. In the domain of communications, there are detailed descriptions of the communication networks connecting internet autonomous systems, connections among peer-to-peer applications, etc. In the biological domain, there are protein interaction networks, metabolic networks, and entire maps of neuronal connections of simple organisms.

In the domain of collaborative production, there are several online collections of works for which it is possible to infer a coauthorship or collaboration network. We have mentioned collaborations among scientists (e.g. the NetHEPT dataset introduced in Section 3.2.2, which is available at http://research.microsoft.com/en-us/people/weic/graphdata.zip), the same applies to actors, dancers, musicians and athletes in team sports, whose information is available from specialized databases. A special case are Wikipedia editors, given that its platform logs almost all the activities of editors. These include co-editing the same article, sending messages to each other through their user profiles, and discussing

about an article on an article discussion page. A number of networks can be inferred from these exchanges.

In general, almost any online platform in which users can declare explicitly their "friends" or connections can provide to a certain extent with information that is relevant to study information propagation: online forums and online games are obvious examples of this.

Some example repositories in which the networks described above can be found include `http://snap.stanford.edu/data/`, `http://www-personal.umich.edu/~mejn/netdata/`, and `http://networkdata.ics.uci.edu/`

6.5 OTHER OFF-LINE DATASETS

Before the information technology revolution, social scientists collected information through direct observation in the field. One of the earliest examples of off-line social networks available is a study of the participation of 18 women in 14 social activities over 9 months in the South of the US [Davis et al., 1941]. A more well-known example is a field work done by Zachary [1977] observing acquiantances between members of a Karate Club of 34 members (dataset available at `http://networkdata.ics.uci.edu/data.php?id=105`). Interestingly, the club splitted in two during the observation period, providing a natural experiment that has been used to benchmark graph partitioning algorithms, e.g. [Girvan and Newman, 2002].

There are many more examples of "off-line" social networks and they are very varied. For instance, the already mentioned dataset of medical records of 12,067 patients during 32 years studied by Christakis and Fowler [2007b], a romantic network of relationships among 288 high-school students obtained by Bearman et al. [2004], and a network of presumed acquaintances links between 74 terrorist suspects analyzed by Krebs [2002].

Multiple datasets of this type can be found in the listing of UCINET IV datasets `http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm`

6.6 PUBLISHING YOUR OWN DATASETS

This is an evolving area and many new datasets are available every year. Publishing a new dataset generates a social good and increases the scientific impact of those who make such datasets available. We briefly outline three steps in a data release: construction, anonymization, and licensing.

First, the construction of each new dataset must be documented in detail. The researcher(s) that release a dataset can not foresee all the different settings in which the dataset will be used. Different settings may require to understand different ways in which the biases of the data must be accounted for. Hence, every sampling, filtering, and processing step must be carefully explained.

Software	License	Language (and mode)	Generate	Operate	Visualize
Gephi	GPL	Java (gui)	Yes	Yes	Yes
Pajek	Free (non-commercial use)	Windows (gui)	Yes	Yes	Yes
SNAP	GPL	C++ (cli)	Yes	Yes	No
Webgraph	GPL	Java (cli)	Yes	Yes	No
Graphviz	GPL	C++ (cli)	No	No	Yes

Second, personally identifiable information must be removed from the data, even if the collection is sampled from publicly-available data.

Third, a license must be chosen. For datasets that do not contain any sensitive or proprietary information, the best is to use a wide disclaimer of warranties and to allow maximum freedom when using the dataset. The Creative Commons Zero license used by the CERN, The British Library, Nature, among many others, serves this purpose. `http://wiki.creativecommons.org/CC0_use_for_data`

For datasets that may contain sensitive or proprietary information, access to the data may be conditioned upon acceptance of a set of terms and conditions through an express agreement. This agreement can specify aspects such as duration, purpose (research, or any purpose), disclaimers of warranties, and item deletion policies (i.e. that the users of the data agree to delete partially or completely the dataset upon request).

6.7 SOFTWARE TOOLS

In this section we overview a few software tools available to support this type of research. We start by giving examples of generic graph mining software, which is widely available and fairly mature; we continue with software that can deal with action traces, which is much more rare. Finally, we outline a few efforts on visualizating viral phenomena.

6.7.1 GRAPH SOFTWARE TOOLS

Tools for creating, manipulating, mining and visualizing large-scale graphs have been available for several years. We focus on a small set of well-established tools that are frequently used by researchers on this area.

Gephi `http://gephi.org` and Pajek `http://pajek.imfm.si/doku.php?id=pajek` can be regarded as *graph processing workbenches*. They implement a number of methods to generate, transform, and visualize graphs. Both provide a graphical user interface (GUI), and thus may be better suited for users without experience using command line interfaces (CLI) interfaces.

In some circumstances e.g., when involving very large graphs or batch processing, using software through a CLI may be advisable.

SNAP http://snap.stanford.edu/snap/ is a set of tools developed to handle social networks data. These tools include many general-purpose graph algorithms including performing clustering, computing measures of centrality, etc.

Webgraph http://webgraph.dsi.unimi.it/ and Graphviz http://graphviz.org/ are specialized tools that focus on particular operations. Webgraph is mainly a graph compression platform to handle huge graphs –it can achieve a surprising 2.89 bits per link when dealing with web links data. Graphviz is a visualization software implementing a number of layout algorithms. Both are used through a command-line interface.

Other tools are available here: https://sites.google.com/site/ucinetsoftware/downloads

6.7.2 PROPAGATION SOFTWARE TOOLS

Software that can deal with information/influence propagation is much more rare than generic graph-processing software.

While many authors may share their code with interested researchers upon request, only a few of them release publicly their code. Among those, we can mention A. Goyal http://www.cs.ubc.ca/~goyal/code-release.php, who provides software for implementing greedy seed selection as well as other algorithms from e.g. Goyal et al. [2011b]; and M. Mathioudakis who provides software implementing the inference of a social network among others described in Mathioudakis et al. [2011] http://queens.db.toronto.edu/~mathiou/spine/

Other software such as the Internet Network Simulator http://isi.edu/nsnam/ns/doc/ can simulate propagation of information through a network (e.g. an Internet worm), but does not include any inference algorithms.

6.7.3 VISUALIZATION

In recent years some demonstrations of visualization of information propagation have emerged.

A first type are software that visualize interactions among pairs of users, e.g. phone calls http://senseable.mit.edu/ or direct messages in Twitter http://www.youtube.com/watch?v=ECqzsom7axQ

A second type is software to visualize a specialized aspect of information propagation in social networks in order to analyze it. In general, these methods visualize the sub-graph induced by a single action as it propagates, as in http://blog.socialflow.com/post/5246404319/breaking-bin-laden-visualizing-the-power-of-a-single. Truthy http://truthy.indiana.edu/ deals with detecting fake political grassroot activism (also known as "astro-turfing"). The visualizations they provide allow analysts to quickly detect

anomalous structures among the users using a certain term or hashtag, such as a densely connected graph or a deviation from the expected for polarizing political topics which is a graph made of two dense subgraphs separable by a small graph cut.

A third type is software that generates visualization for consumption by the general public. Mass media is becoming increasingly interested in mining social media as a way of understanding society and/or contributing to the journalistic process. The Guardian, for instance, presented a visualization of rumours as they appear, propagate, and die: `http://www.guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter`. The New York Times' "Project Cascade" depicts the propagation of links to news using a solar system metaphore in which the original news item is the Sun: `http://nytlabs.com/projects/cascade.html`

6.8 CONCLUSIONS

The data used by most researchers in this area tend to suffer from one of the following drawbacks: (i) it does not describe explicitly the action traces, (ii) it does not include a explicit social network, or (iii) it is proprietary.

To a large extent, researchers resort to using synthetic propagations on existing social networking data. While this may be enough for some purposes, it is clearly desirable that experimental validation is done over datasets that resemble closely real-world datasets, specially for research that claims to be significant for real-world applications. Results on proprietary data sources, which may be more realistic, are not reproducible and hence may not be a solid ground for further research.

The lack of appropriate data can seriously slow down research on influence and information propagation. The scarcity of reference implementations of well-known algorithms may hamper efforts to reproduce and compare different algorithms. Both problems can be dealt with if the research community embraces reproducibility as a fundamental value, and collectively shares the datasets and software needed to move forward.