An Analysis of Activities in Facebook

Khanh Nguyen Duc A. Tran
Department of Computer Science
University of Massachusetts - Boston
Email: {knguyen, duc}@cs.umb.edu

***Abstract*—Human dynamics in real-world social networks has been a long lasting topic of research. Recently, the rapid growing of online social networks with their data made accessible has given researchers a door of opportunity to study large-scale social networks. Facebook has been the largest, most influential, and fastest growing social network on the Internet to date. As Facebook's userbase is extremely diverse, spreading across different countries, races, ages, professions, and interests, to name a few, this network is a good model to study our society. In this paper, we present an analysis of a regional Facebook dataset, with attention given to the correlation among user connectivity, activity, and similarity.**

I. Introduction

The study of complex networks started with Erdos and Renyi's random graph model [1]. In this model (ER), two nodes are randomly connected with a fixed probability. The degree distribution of a random graph so built can be characterized by a Poison distribution in which a majority of the nodes have approximately the same degree that is close to the average degree of the nodes. With the advances in computing technologies over the last decade, it has become possible to map the topology of many large real-world networks, including the WWW [2], the Internet [3], metabolic networks [4], protein networks [5], co-authorship networks [6], and sexual contact networks [7]. Interestingly, the topologies of these real-world networks have shown a significant deviation from the ER model. Instead of following Poison, the degree distribution exhibits a power-law tail; i.e., for large degree k, the fraction of nodes having this degree is $P(k) \propto k^{-\lambda}$. The parameter λ, called the power-law exponent for the degree distribution.

This paper is focused on online social networks, a special kind of complex networks that has received a lot of attention recently. An online social network is a social network that is formed by users of the Internet. As defined in [8], a social network is "a set of people or groups of people with some pattern of contacts or interactions between them". Starting a long time ago with Migram's experiment, the importance of understanding foundational mechanisms behind social networks has spurred many interesting research problems, especially with the emergence of today's online social networks such as Facebook, MySpace, LinkedIn, Flickr, and Youtube. People have found these networks a convenient way to socially interact with each other. The popularity of these networks, of which a lot of data has been made available, provides researchers with a great opportunity to examine and infer sociological theories and implications by analyzing the topological structures of the network and the patterns of interaction between the users. This gives us insights into what underlines human dynamics, helping to answer questions such as what is the key factor that makes a person popular in a network, or how active is a relationship between a popular person and a less popular person. Social scientists can validate or invalidate traditionally-found theories about offline social networks by utilizing the data available from online networks. Furthermore, as online social networks account for a large portion of current Internet traffic, understanding their structure and growth is useful to how we shape the Internet in the future [9].

In this paper, we present an analysis of Facebook using the dataset from [10]. While a comprehensive study of this network has been reported in [10], we have conducted our study in a different lens. Specifically, we attempt to answer the following questions:

- Contact network versus activity network: are they the same or different? is it true that a user with many contacts also participates in many activities? does a user communicate with all of its contacts as often, or with just a few of them?
- Activity distribution: how activities are distributed among users? is there a certain set of users who participate most actively in the network?
- User similarity versus activity: can we infer user similarity from activity and vice versa? do most activities take place between "similar" people or "different" people? between a popular person and a less popular person?

The remainder of this paper is organized as follows. We discuss the related work in Section II. We describe the dataset and the previous findings of [10] in Section III. We present the results of our analysis in Section IV. The paper is concluded in Section V with pointers to our future work.

II. Related Work

Barabasi and Albert suggested in [11] that the topological properties of many real-world networks can be explained through a dynamic process in which a network grows in time steps based on a preferential attachment mechanism, rather than a static process previously presumed. Since its inception, the preferential attachment model has become the basis for many subsequent generative power-law models. In these models, the preference of choosing a node to link to is solely a non-decreasing function of this node's degree. As a result, older nodes will increasingly have more edges. Other work has been devoted to validate this model or suggest other alternatives for real-world networks [**?**], [12], [13].

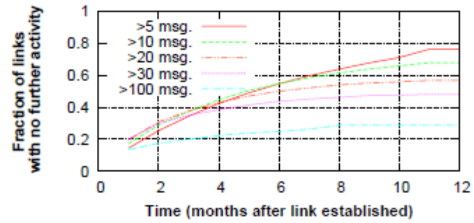
An early paper on online social networks is [14] which studies the Club Nexus website of Stanford Universty. Yahoo 360 (now deceased) is analized in [15]. A series of work on popular online social networks, namely Orkut, Flickr, Youtube and Facebook, is reported in [10], [16], [17] [9]. The general finding is that all these networks exhibit small-world and scale-free properties, somewhat consistent with the preferential attachment model [18]. The evolution of activity inside a social network has also been investigated. For example, the activity network of Facebook and Cyworld social networks is analyzed in [10] and [19], respectively. In a different effort, the authors of [20] study user communication patterns in an Instant Messenger network, using a dataset of more than 300 billion messages. It is found that a strong homophily exists with the tendency that similar users tend to interact with each other (based on location, sex, age, etc.).

III. The Dataset

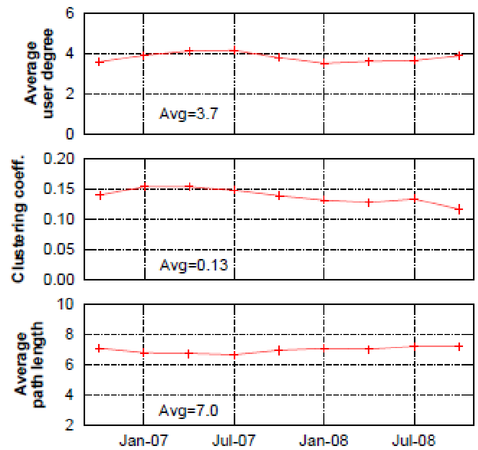
Facebook is the largest online social network to date. It boasts 300 million users worldwide. According to iStrategyLabs.com, Facebook's US user base jumped 150% from 2009 to 2010, reaching more than 100 million users. Facebook allows a user to create a public profile with pictures and other personal information such as gendre, date of birth, hometown, phone number, school, employer, interests, and current GPS-location. Each user has a list of friends, but no more than 5000 of them. Two users can establish a friendship link by sending and accepting a friendship request. Each user has a "wall" wherein his or her friends can write a message. A wall message is visible to everyone who has access to the user's profile. Posting wall messages is the mostly used activity in Facebook. Hereafter, we call the network of friendship links the *contact network*, and that of wall-posting links the *activity network*.

We use the data made available by Max-Planck Software Institute for Software Systems [21]. There are two separate datasets, each collected by a crawler program during a different time period. The details of how these crawlers work can be found in [10]. The first crawl began on 12/28/2008 and ended on 1/3/2009, recording only the friendship links (together with their timestamps) which form the contact network. As a result, we have a network of 90,269 users (52% in New Orleans) and 3,646,662 friendship links. The second crawl was focused on the wall-posting information of the users who have been reached in the first crawl. This crawl was conducted between 1/20/2009 and 1/22/2009. The timestamp of each wall message and the identifiers of its sender and receiver were recorded. For this second dataset, we have a network of 60,290 users, 188,892 links, and 838,092 wall messages posted.

Looking at the data as a time series to observe the evolution of the activity network over time, the authors of [10] have found that the activity pattern (i.e., regarding user wall posting) changes significantly over time. For example, most messages between two users happens very early after they start writing to each other; afterwards, their interaction becomes rapidly less frequent. Regardlessly, many of the network's topological



(a) Fraction of user pairs with no further activity after a given number of months



(b) Properties of the evolving activity network in terms of three popular metrics

Fig. 1. Previous findings [10]

properties (e.g. average node degree, average path length, clustering coefficient, etc.) remain stable. These findings are ilustrated in Figure 1(a) and Figure 1(b).

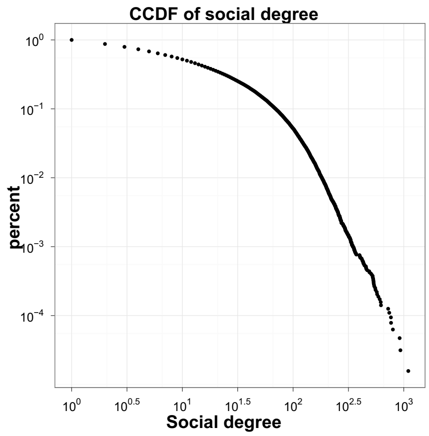
IV. Our Analysis

The goal of our study is to answer the questions raised in Section I. We use R [22] as the programming environment for our analysis and employ two packages: igraph [23] for network analysis and ggplot2 [24] for plotting the results. We use the algorithm of [25] for power-law fitting.

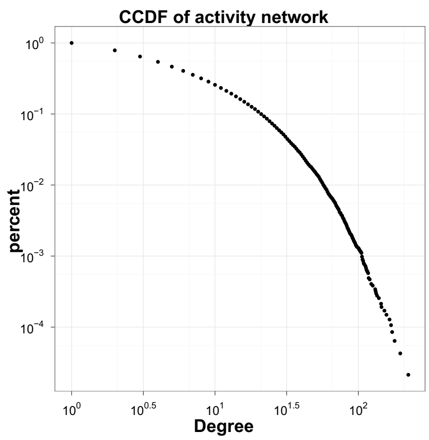
A. Degree Distribution: Contact Network and Activity Network

First, we look at the degree distribution of the contact network. Here, a node's degree is the number of its friends. As suggested by the literature of power-law networks, we expected the degree distribution of large-scale online social networks to follow a power-law distribution. Figure 2(a) shows the CCDF (complementary CDF) of the degree distribution of the friendship dataset. It is shown that for degrees above *xmin* = 96, the degree distribution is power-law with exponent 3.5 (error $D = 0.035$).

Figure 2(b) shows the degree distribution from the activity network constructed from the wall-message dataset. In this network, a link is established between two users if one user



(a) Contact network



(b) Activity network

Fig. 2. Degree distribution: Weak power law is observed for both contact network and activity network.

writes on the other user's wall. In Facebook, to write on someone's wall, an user must be in his or her friendlist. Hence, the activity network is a subgraph of the friendship network. The activity network also exhibits a power-law degree distribution for its tail ($xmin$ = 34), approximately with the same exponent (3.5) with that of the contact network.

B. Node Activity: Distribution and Relationship with Node Connectivity

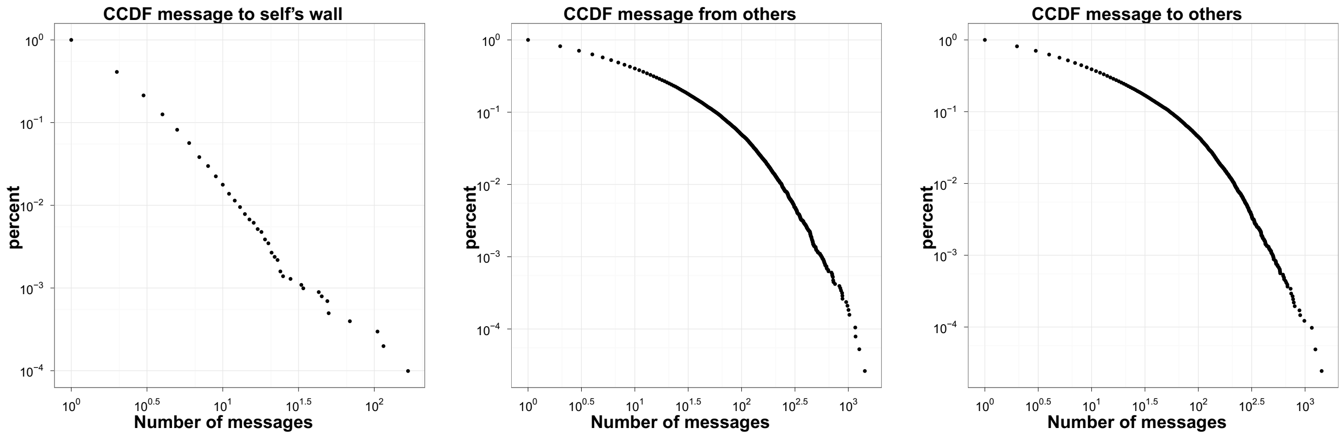
To further explore the presence of power law in the activity network, we look at the message count per user in three cases. In the first case, we are interested in each user's activity in posting to his or her own wall. Self-posting occurs to either respond to a message from other users, or to post an announcement for friends to know. In the second case, we count the number of messages a user receives from his or her friends. In the third case, we count the number of messages a user writes to other walls. The activity of a user is largely dominated by message posting to other walls. The activity distributions for these cases are demonstrated in Figure 3(a), Figure 3(b), and Figure 3(c), respectively. Power-law is quite obvious in the distribution of the number of self-posted messages (exponent 3.19), but it is less obvious in the case of received messages and the case of messages posted to other walls. In the latter two cases, the power-law appears late in the tail with exponent 3.5 for both.

We investigate the correlation between a node's level of activity and its degree in the activity network (i.e., the number of friends whom an user interacts with). We plot the message count for each node, sorted in the non-decreasing order of degree, in Figure 4. One might think that a user with a few friends would be less active than a user with more friends. As shown in Figure 4(b) and Figure 4(b), this is observed for the case where activity is defined to be "writing messages to others or receiving messages from others". However, it is not true for the case where activity is defined to be "posting messages to own wall". Indeed, Figure 4(a) shows that there are nodes that have large degree (e.g., larger than 200) yet posting very few messages to their own wall, and there are nodes that have very small degree (e.g., less than 20) yet posting a lot of messages to their own wall. It would be interesting to identify "who" these people are in the real world.

C. Link Activity: Distribution and Relationship with Similarity

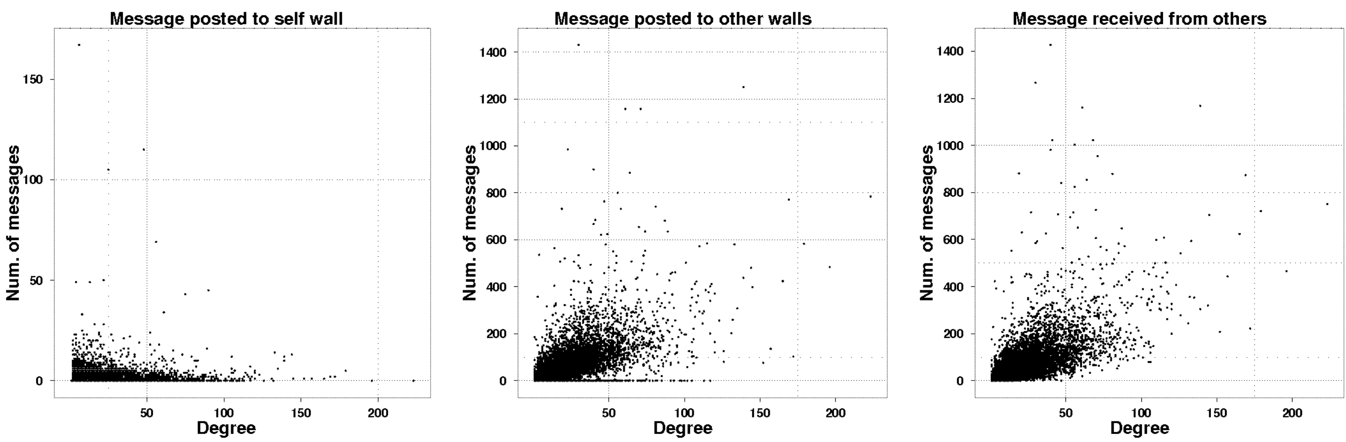
Next, we count the number of messages between every pair of users in the activity network. In the *directed* case (Figure 5(a)), for each directed link A-to-B we count the number of messages that A writes on B's wall and, similarly, for directed link B-to-A we count the number of messages that B writes on A's wall. Links A-to-B and B-to-A are considered different links. In the *undirected* case (Figure 5(b)), we treat these two links as the same link A-B, counting the total number of messages between A and B. The message-posting activity distributions in both cases show strong inclination toward power law, with exponent 3.34 in the directed case and 2.59 in the undirected case. The activity inside the network is dominated largely by a few pair of users who exchange hundreds of messages while the rest seldomly interact.

To study the pattern of interaction regarding who often communicates with whom, whether message posting occurs mostly between people sharing some similarity or no similarity, we plot Figure 6 the message count per link, sorted increasingly according to the difference between the link's end nodes' degrees. As observed in Figure 6(a), the message-count distribution among all the links having the same degree difference follows a pattern which is similar regardingless of this degree difference. The pattern suggests a power-law distribution; e.g., Figure 6(b) shows the plot of the message counts of links degree difference of 1, which clearly shows a power-law pattern (exponent 2.57). In other words, whether the degree difference is 0, or 1, or 2, etc., there are a few links of this degree difference that post a majority of messages while the remaining links of this degree difference are mostly idle. The implication here is two-fold. First, the degree difference is not a significant factor to determine the level of interaction between users. An user with 100 friends is as likely to interact



(a) Messages to own wall (b) Messages received from others (c) Messages to other walls

Fig. 3. Message count per user: distribution



(a) Messages to own wall, sorted by degree (b) Messages from others, sorted by degree (c) Messages to other walls, sorted by degree

Fig. 4. Message count per user, sorted by degree

with an user with 10 friends as much as with another user with 100 friends. Second, the dominance of a small set of users on the activity of the network is not only present globally but also *locally* when we divide the network into groups based on degree difference.

V. CONCLUSIONS

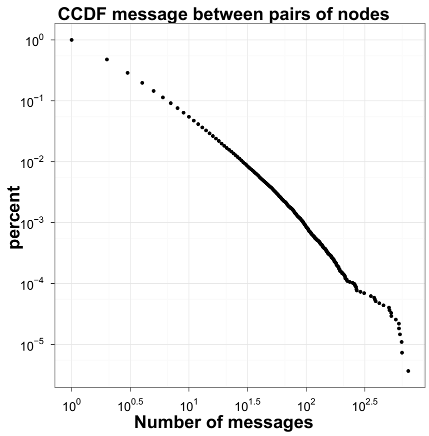
In analysis of wall-posting activities using a regional Facebook dataset, we have found that power-law is observed, albeit weakly, in the distributions of node degree in both contact and activity networks and of the number of messages per node (self-posted, received, and sent to other walls). A stronger presence of the power-law pattern is observed in the distribution of per-link activity in the activity network; i.e., in terms of the number of messages exchanged on a link of two nodes. Also, some interesting findings are worth further investigation. There are complex networks for which we know that there is a tendency for a node to connect with another of a similar degree (e.g., traditional social networks) or of a different degree (e.g., technological networks). However, this tendency is not clear in Facebook; the intensity of message posting involving two users does not depend clearly on their degree similarity or difference. Secondly, we have found a correlation between an user's popularity and activity based on the number of messages posted to other walls and received from other users, but not based on how often he or she writes to own wall. In future work, we will investigate these issues further and also look at other on-line social networks to see if similar activity patterns are present.

ACKNOWLEDGMENT

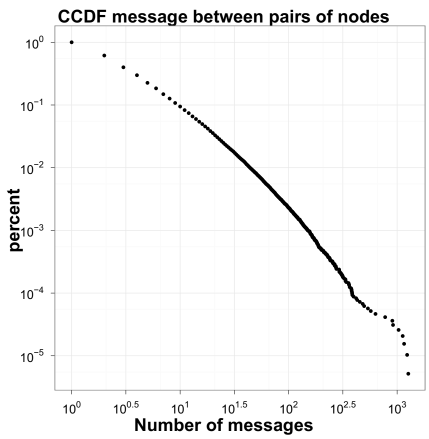
The authors would like to thank the Max-Planck Institute for Software Systems, especially Alan Mislove, for sharing the dataset that was used for this study.

REFERENCES

[1] P. Erdós and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

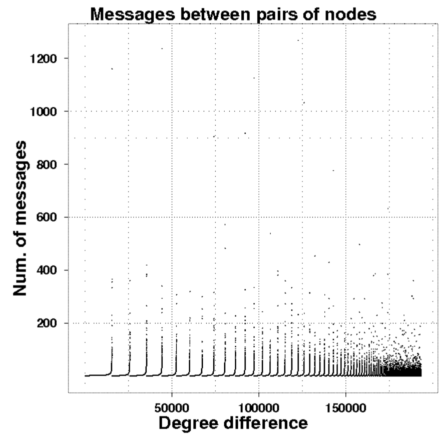


(a) Directed

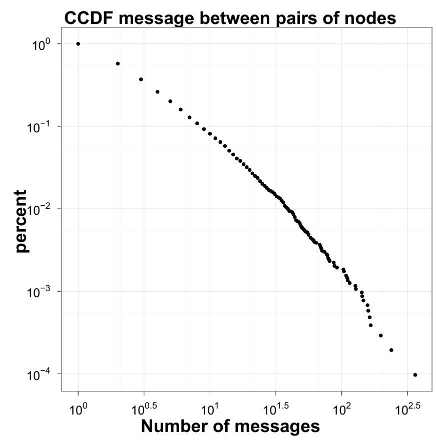


(b) Undirected

Fig. 5. Message count per link: Power-law distribution is clearly observed.



(a) Increasingly sorted by degree difference



(b) Message count with degree difference 1

Fig. 6. Message count per link: degree similarity vs. activity.

[2] R. Albert, H. Jeong, and A. L. Barabási, "Diameter of the world-wide-web," *Nature*, vol. 400, pp. 107–110, 1999.

[3] M. Faloutros, P.Faloutros, and C. Faloutros, "On power-law relationship of the internet topology," *ACM SIGCOM 99, Comp. Comm. Rev.*, vol. 29, pp. 251–260, 1999.

[4] H. Jeong, B. Tombor, R. Albert, N. Oltvai, and A. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 607, p. 651, 2000.

[5] H. Jeong, S. Mason, R. Oltvai, and A. Barabási, "Lethality and centrality in protein networks," *Nature*, vol. 411, p. 41, 2001.

[6] M. Neuman, "The structure of scientific collaboration networks," *Proc. Nat. Acad. Sci. USA*, vol. 98, pp. 404–409, 2001.

[7] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Aberg, "The web of human sexual contacts," *Nature*, vol. 411, p. 907, 2001.

[8] M. Newman, "The structure and function of complex networks," *SIAM Review*, 2003.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," *Proceedings of the 5th ACM/USENIX Internet Measurement Conference*, 2007.

[10] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks*, 2009.

[11] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[12] C. Bedogne and G. J. Rodgers, "Complex growing networks with intrinsic vertex fitness," *Phys. Rev. E*, vol. 74, p. 046115, 2006.

[13] J. S. Kong, N. Sarshar, and V. P. Roychowdhury, "Experience versus talent shapes the structure of the web," *CoRR*, vol. abs/0901.0296, 2009.

[14] L. A. Adamic, O. Buyukkokten, and E. Adar, "A social network caught in the web," *First Monday*, 2003.

[15] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," *In Proc. ofKDD*, 2006.

[16] A. Mislove, K. P. Gummadi, and P. Druschel, "Exploiting social networks for internet search," *In Proc. of HotNets*, 2006.

[17] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network." *In Proc. of WOSN*, 2008.

[18] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science, vol 286, no 5439*, 1999.

[19] H. Chun, H. Kwak, Y. Eom, Y. Ahn, S. Moon, and H. Jeong, "Comparison of online social relations in volumevs interaction: a case study of cyworld," *In Proc. of Internet Measurement Conference*, 2008.

[20] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," *In Proc. of the International World Wide Web Conference, Beijing, China*, 2008.

[21] MPI-SWS, http://socialnetworks.mpi-sws.org/.

[22] R. Project, http://www.r-project.org/.

[23] iGraph, http://igraph.sourceforge.net/.

[24] ggplot2, http://had.co.nz/ggplot2/.

[25] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.