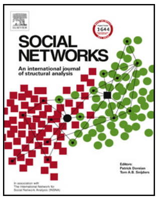
Social Networks 41 (2015) 48–55

Contents lists available at ScienceDirect

Social Networks

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



# Algorithms for diversity and clustering in social networks through dot product graphs☆

Matthew Johnson[a], Daniël Paulusma[a,*], Erik Jan van Leeuwen[b]

[a] *School of Engineering and Computer Science, Durham University, United Kingdom*
[b] *Max-Planck Institut für Informatik, Saarbrücken, Germany*

ARTICLE INFO

*Keywords:*
Social network
*d*-Dot product graph
Independent set
Clique

ABSTRACT

In this paper, we investigate a graph-theoretical model of social networks. The *dot product model* assumes that two individuals are connected in the social network if their attributes or opinions are similar. In the model, a $d$-dimensional vector $\mathbf{a}^v$ represents the extent to which individual $v$ has each of a set of $d$ attributes or opinions. Then two individuals $u$ and $v$ are assumed to be friends, that is, they are connected in the graph model, if and only if $\mathbf{a}^u \cdot \mathbf{a}^v \geq t$, for some fixed, positive threshold $t$. The resulting graph is called a *d-dot product graph.*

We consider diversity and clustering in social networks by using a $d$-dot product graph model for the network. Diversity is considered through the size of the largest independent set of the graph, and clustering through the size of the largest clique. We present both positive and negative results on the potential of this model. We obtain a tight result for the diversity problem, namely that it is polynomial-time solvable for $d=2$, but NP-hard for $d \geq 3$. We show that the clustering problem is polynomial-time solvable for $d=2$. To our knowledge, these results are also the first on the computational complexity of combinatorial optimization problems on dot product graphs. We also give new insights into the structure of dot product graphs.

We also consider the situation when two individuals $u$ and $v$ are connected if and only if their preferences are not antithetical, that is, if and only if $\mathbf{a}^u \cdot \mathbf{a}^v \geq 0$, and the situation when two individuals $u$ and $v$ are connected if and only if their preferences are neither antithetical nor "orthogonal", that is, if and only if $\mathbf{a}^u \cdot \mathbf{a}^v > 0$. For these two cases we prove that the diversity problem is polynomial-time solvable for any fixed $d$ and that the clustering problem is polynomial-time solvable for $d \leq 3$.

## 1. Introduction

Social networks are often modeled by a graph in order to use advanced algorithmic (or statistical) tools. Indeed, there is a large body of literature on (random) graph models for social networks (see, for example, the surveys by Newman (2003) and Snijders (2011)). These studies have proposed many models for social networks, offering different explanations of why connections are made in the network (see the partial overview in Liben-Nowell and Kleinberg (2003)). For example, the models of Simon (1955), de Price (1976), and Barabási and Albert (1999) famously propose that if you have many friends, you are more likely to make further new friends. A similar idea was recently considered from an algorithmic perspective by Bhawalkar et al. (2012).

We consider a different predictor for connections in a social network, namely the degree of similarity of attributes and opinions of different individuals. Generally, individuals with similar attributes or opinions are more likely to be connected. This is known as the *homophily principle* and is well-studied in sociological research (see, for example, the survey by McPherson et al. (2001)). To model the attributes of an individual $u$, we can associate them with a vector $\mathbf{a}^u$, where an entry $a_i^u$ expresses the extent to which $u$ has an attribute or opinion $i$ (Watts et al., 2002). For example, a positive value of $a_i^u$ could indicate that $u$ likes item $i$, whereas a negative value suggests that $u$ dislikes item $i$. We call this a *vector model.*

There are many ways to measure similarity using a vector model (see, for example, Adamic and Adar, 2003; Hoff et al., 2002; Kim and Leskovec, 2010; Leskovec et al., 2010; Watts et al., 2002). We will use the dot product as a similarity measure. This measure is closely related to the cosine measure, which was studied before by

☆ This paper was supported by EPSRC (EP/G043434/1), and an extended abstract of this paper has appeared in the Proceedings of ISAAC 2013 (Johnson et al., 2013).

* Corresponding author. Tel.: +44 0191 33 41723; fax: +44 0191 33 41701.
*E-mail addresses:* matthew.johnson2@durham.ac.uk (M. Johnson), daniel.paulusma@durham.ac.uk (D. Paulusma), erikjan@mpi-inf.mpg.de (E.J. van Leeuwen).

http://dx.doi.org/10.1016/j.socnet.2015.01.001
0378-8733/© 2015 Elsevier B.V. All rights reserved.

researchers in information retrieval and social networks (see e.g. Cattuto et al., 2008; Crandall et al., 2008). The dot product measure leads to the *dot product model* for social networks, which is defined as follows. Consider a social network that consists of a set $V$ of individuals, together with a vector model $\{\mathbf{a}^u \mid u \in V\}$. Let

$$\mathrm{sim}(u, v) = \mathbf{a}^u \cdot \mathbf{a}^v = \sum_{i=1}^{d} a_i^u a_i^v.$$

If the similarity $\mathrm{sim}(u, v)$ is at least some specified *threshold* $t > 0$, then we view the preferences of $u$ and $v$ to be sufficiently close together for $u$ and $v$ to be connected, that is, to be friends within the network. This immediately implies a graph $G = (V, E)$, where $(u, v) \in E$ if and only if $\mathrm{sim}(u, v) \geq t$. Such a graph is called a *dot product graph* of *dimension d*, or a *d-dot product graph*. The vector model $\{\mathbf{a}^u \mid u \in V\}$ together with the threshold $t$ is called a *d-dot product representation* of $G$.

The dot product model has a long tradition, both in the study of social networks (see, for example, Breiger, 1974) and in (algorithmic) graph theory (see, for example, Reiterman et al., 1989a,b, 1992 and particularly Fiduccia et al., 1998). Below, we survey some of the recent work and how it relates to social networks.

The dot product graph as a model for social networks was formalized by Nickel (2007), Young and Scheinerman (2007, 2008), Minton (2008), and Scheinerman and Tucker (2010). In particular, these works consider a randomized version of the dot product model, where the dot product of two vectors constitutes the probability that an edge occurs between the corresponding vertices. This randomized version of the model fits in a long line of research on random graph models for social networks, such as the classic Erdös–Rényi graph model (Erdös and Rényi, 1959), the Kronecker graph model (Leskovec et al., 2010) and the multiplicative attribute graphs model (Kim and Leskovec, 2010) (which generalizes the Kronecker graph model). The random dot product graph model exhibits the main characteristics that one would expect from a model for social networks, such as the property that two vertices are more likely to be adjacent if they have a common neighbor, the small-world principle, and a power-law degree distribution (Nickel, 2007). Studies into the dot product model were also motivated by the work of Papadimitriou et al. (2000) and Caldarelli et al. (2008). Moreover, dot product graphs share some ideas with low-complexity graphs (Arora et al., 2009).

Dot product graphs have been studied from the perspective of (algorithmic) graph theory mostly with respect to the question of determining the *dot product dimension* of a graph: the minimum dimension $d$ for which a graph has a $d$-dot product representation. This notion is well defined, as every graph on $m$ edges has a dot product representation of dimension $m$ (Fiduccia et al., 1998). Observe that, in the context of social networks, the dot product dimension can be seen as the smallest number of preferences needed to determine all friendship relations and non-relations between any two individuals in the network. Hence, the dot product dimension is a measure of the social complexity of a network (Minton, 2008).

The work of Fiduccia et al. (1998) implies that deciding whether a graph has dot product dimension 1 takes polynomial time. However, Kang and Müller (2012) showed the problem of deciding whether a graph has dot product dimension $d$ is NP-hard for all fixed $d \geq 2$ (membership in NP is still open). They also proved that an exponential number of bits is sufficient and can be necessary to store a $d$-dot product representation of a dot product graph. Kang et al. (2011) gave a tight bound of 4 on the dot product dimension of a planar graph. Fiduccia et al. (1998) conjectured that any graph on $n$ vertices has dot product dimension at most $n/2$; Li and Chang (2014) recently confirmed this conjecture for a number of graph classes.

In this paper, we study how the complexity of computing structural properties of a social network is influenced by the complexity of the network's dot product model. Note that many standard structural properties, such as the graph diameter and the clustering coefficient, are easy to compute even on general graphs. Therefore, we consider two more advanced structural properties that give information on diversity and clustering in the network. These properties relate to classic graph optimization problems that are NP-hard to compute on general graphs, but whose computational complexity on dot product graphs was unknown. In fact, to the best of our knowledge (see also Spinrad, 2003, p. 309), no algorithmic work on graph optimization problems on $d$-dot product graphs for $d \geq 2$ has been done prior to this work.

The main observation from our study is that when computing information on diversity and clustering properties of a social network, it is helpful if the network has small dot product dimension. When the network has small dot product dimension, we give positive results, in the sense of polynomial-time algorithms, for the studied problems. When the network does not have small dot product dimension, we observe clear barriers that prevent us from generalizing our algorithms. Additionally, we give a hardness result for one of the problems. This furthers our understanding of the scope of this particular model for social networks, but more importantly suggests that future studies on dot product models should focus on investigating approximation or fixed-parameter algorithms for the studied problems.

This main observation is supported by the following results. First, we consider diversity, by finding (the size of) a largest group of individuals in the network that are different-minded, and thus pairwise disconnected. This corresponds to the well-known Independent Set problem, which is NP-hard on general graphs (Karp, 1972). On 1-dot product graphs the problem is known to be solvable in polynomial time, since such graphs consist of at most two connected components, each of which is a threshold graph (Fiduccia et al., 1998), and Independent Set has a trivial polynomial-time algorithm for threshold graphs.[1] However, its complexity on $d$-dot product graphs for $d \geq 2$ is open. We settle this by proving that Independent Set is polynomial-time solvable on 2-dot product graphs, but becomes NP-hard on 3-dot product graphs.

Second, we consider clustering, by finding (the size of) a largest group of individuals in the network that are like-minded, and thus pairwise connected. This corresponds to the well-known Clique problem, which is NP-hard on general graphs (Karp, 1972). Again, on 1-dot product graphs a trivial polynomial-time algorithm is known using the relation to threshold graphs (Fiduccia et al., 1998), but its complexity has not been analyzed on $d$-dot product graphs for $d \geq 2$. We give initial insights into the complexity of this problem and show that it is polynomial-time solvable on 2-dot product graphs.

We remark that our complexity results depend on a number of lemmas on the structure of dot product graphs which are of independent interest.

To complement these results, we consider two variants of the dot product model. For the first variant, we model the scenario in which two individuals are connected if their preferences are not antithetical. That is, consider the graph where two individuals $u, v$ are connected if and only if $\mathbf{a}^u \cdot \mathbf{a}^v \geq 0$. We call such a graph a $d^0$-*dot*

[1] A possible definition of a threshold graph states that $G$ is a threshold graph if it can be constructed from a single vertex by repeatedly adding an isolated vertex or a dominating vertex (that is, a vertex adjacent to all other vertices) (Golumbic and Trenk, 2004; Mahadev and Peled, 1995). Using this definition, a polynomial-time algorithm for Independent Set (and for Clique) can be easily derived.