

## Durham Research Online

---

### Deposited in DRO:

15 January 2015

### Version of attached file:

Accepted Version

### Peer-review status of attached file:

Peer-reviewed

### Citation for published item:

Johnson, M. and Paulusma, D. and van Leeuwen, E.J. (2013) 'Algorithms to measure diversity and clustering in social networks through dot product graphs.', in Algorithms and computation : 24th International Symposium, ISAAC 2013, Hong Kong, China, 16-18 December 2013 ; proceedings. Berlin, Heidelberg: Springer, pp. 130-140. Lecture notes in computer science., 8283

### Further information on publisher's website:

[http://dx.doi.org/10.1007/978-3-642-45030-3\\_13](http://dx.doi.org/10.1007/978-3-642-45030-3_13)

### Publisher's copyright statement:

The final publication is available at Springer via [http://dx.doi.org/10.1007/978-3-642-45030-3\\_13](http://dx.doi.org/10.1007/978-3-642-45030-3_13)

### Additional information:

## Use policy

---

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full DRO policy](#) for further details.

# Algorithms to Measure Diversity and Clustering in Social Networks through Dot Product Graphs

Matthew Johnson<sup>1</sup>, Daniël Paulusma<sup>1</sup>, and Erik Jan van Leeuwen<sup>2</sup>

<sup>1</sup> School of Engineering and Computer Science, Durham University, England  
{matthew.johnson2,daniel.paulusma}@durham.ac.uk

<sup>2</sup> Max-Planck Institut für Informatik, Saarbrücken, Germany  
erikjan@mpi-inf.mpg.de

**Abstract.** Social networks are often analyzed through a graph model of the network. 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 two measures for diversity and clustering in social networks by using a  $d$ -dot product graph model for the network. Diversity is measured through the size of the largest independent set of the graph, and clustering is measured through the size of the largest clique. We obtain a tight result for the diversity problem, namely that it is polynomial-time solvable for  $d = 2$ , but NP-complete 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 consider the situation when two individuals are connected if their preferences are not opposite. This leads to a variant of the standard dot product graph model by taking the threshold  $t$  to be zero. We prove in this case that the diversity problem is polynomial-time solvable for any fixed  $d$ .

## 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 [23] and Snijders [32]). Many of these studies verify that a particular model has properties that have been observed in real-world social networks, such as a power-law degree distribution or the small-world principle, but do not consider why connections are made in the first place. This has led to the development of models that do take such reasons into account (a partial overview is in Liben-Nowell and Kleinberg [21]). For example, the models of Simon [31], Price [26], and Barabási and Albert [3] famously pose that if you have many friends, you are more likely

to befriend more people. A similar type of engagement was recently considered from an algorithmic perspective by Bhawalkar et al. [6].

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 has a long tradition within sociological research (see, for example, the survey by McPherson et al. [22]). 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  $i$  [33]. 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, [1, 14, 19, 33]). We will use the dot product as a similarity measure, leading to the *dot product model* for social networks. Formally, this model 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

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

If the similarity  $\text{sim}(u, v)$  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  $\text{sim}(u, v) \geq t$ . 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 graph as a model for social networks was recently formalized by Nickel, Scheinerman, Tucker, and Young [24, 30, 34, 35]. Their studies were motivated by earlier work of Papadimitriou et al. [25] and Caldarelli et al. [7]. However, dot product graphs have a much longer tradition, both in sociology (see, for example, Breiger [5]) and in graph theory. We briefly survey known graph-theoretic results. Reiterman et al. [27–29] and particularly Fiduccia et al. [10] proved several structural results. The work of Fiduccia et al. [10] implies that 1-dot product graphs can be recognized in polynomial time. However, Kang and Müller [17] showed the problem of deciding whether a graph has dot product dimension  $d$  is NP-hard for all fixed  $d \geq 2$  (membership of 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. There are several papers that consider the minimum dimension  $d$  such that a graph is a  $d$ -dot product graph (the *dot product dimension* of a graph) [16, 20], deriving for example a tight bound of 4 on the dot product dimension of a planar graph [16]. Finally, dot product graphs share some ideas with low-complexity graphs [2].

In this paper, we consider the complexity of computing advanced structural measures of social networks through the dot product model. Note that many standard structural measures, such as the graph diameter and the clustering co-

efficient, are easy to compute. Therefore, we consider two more advanced measures for diversity and clustering. These are related to classic graph optimization problems whose computational complexity on dot product graphs was unknown. In fact, to the best of our knowledge, our work provides the first complexity results for graph optimization problems on dot product graphs.

First, we consider a measure for 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-complete, W[1]-complete, and very hard to approximate on general graphs [18, 9, 13], but its complexity on dot product graphs is open. We settle this by proving that INDEPENDENT SET is polynomial-time solvable on 2-dot product graphs, but becomes NP-complete on 3-dot product graphs.

Second, we consider a measure for 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 also NP-complete, W[1]-complete, and very hard to approximate on general graphs [18, 9, 13], but its complexity has not been analyzed on dot product graphs. We give initial insights into the complexity of this problem and show that it is polynomial-time solvable on 2-dot product graphs.

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 opposite. 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 product graph. Recall that in  $d$ -dot product graphs, the threshold  $t$  for connectivity must be greater than zero, and hence the definition of  $d^0$ -dot product graphs is different. Moreover, the structure of  $d^0$ -dot product graphs is substantially different from that of  $d$ -dot product graphs. To illustrate this, we prove that INDEPENDENT SET is polynomial-time solvable on  $d^0$ -dot product graphs for any fixed  $d$  and that CLIQUE is polynomial-time solvable if  $d \leq 3$ .

For the second variant, we model the situation in which two individuals are connected in the model if their preferences are neither opposite nor orthogonal. Consider the graph that is obtained when two vertices  $u, v$  are adjacent if and only if  $\mathbf{a}^u \cdot \mathbf{a}^v > 0$ . We call this a  $d^+$ -dot product graph. It follows from Fiduccia et al. [10] that the graph class where two vertices are adjacent if and only if  $\mathbf{a}^u \cdot \mathbf{a}^v > t$  for some  $t > 0$  is equivalent to the class of  $d$ -dot product graphs. However, we prove that the structure of  $d^+$ -dot product graphs is different from that of  $d$ -dot product graphs and that of  $d^0$ -dot product graphs. Still, we can show that INDEPENDENT SET is polynomial-time solvable on  $d^0$ -dot product graphs for any fixed  $d$ , as is CLIQUE when  $d \leq 3$ .

We provide an overview of our results in Table 1.

## 2 Preliminaries

All graphs that we consider are finite, undirected, and have neither loops nor multiple edges. For undefined graph terminology we refer to Diestel [8].

Setting	INDEPENDENT SET	CLIQUE
$d$ -DPG ( $\geq 1$ )	in P for $d \leq 2$ NP-complete for $d \geq 3$	in P for $d \leq 2$ ? for $d \geq 3$
$d^0$ -DPG ( $\geq 0$ )	in P for $d \geq 0$	in P for $d \leq 3$ ? for $d \geq 4$
$d^+$ -DPG ( $> 0$ )	in P for $d \geq 0$	in P for $d \leq 3$ ? for $d \geq 4$

**Table 1.** An overview of our results for the problems INDEPENDENT SET and CLIQUE on  $d$ -dot product graphs (the first row),  $d^0$ -dot product graphs (the second row), and  $d^+$ -dot product graphs (the third row), respectively, for fixed dimension  $d$ .

Let  $G = (V, E)$  be a graph. We denote the neighbourhood of a vertex  $u \in V$  by  $N(u) = \{v \mid (u, v) \in E\}$ . A subset  $U \subseteq V$  is *independent* if no two vertices in  $U$  are joined by an edge, and  $U$  is a *clique* if every two vertices of  $U$  are adjacent. Given  $U \subseteq V$ ,  $G[U]$  denotes the subgraph of  $G$  induced by  $U$ , that is, it has vertex set  $U$  and an edge between two vertices of  $U$  if and only if  $G$  has an edge between them. The *complement* of  $G$  has vertex set  $V$  and an edge between two distinct vertices if and only if these vertices are not adjacent in  $G$ .

A graph is a *comparability graph* if there exists an assignment of exactly one direction to each of its edges such that  $(a, c)$  is a directed edge whenever  $(a, b)$  and  $(b, c)$  are directed edges. The complement of a comparability graph is called a *co-comparability graph*. A graph is  *$p$ -partite* if its vertex set can be partitioned into at most  $p$  independent sets. If  $p = 2$ , then the graph is called *bipartite*. The complement of a  $p$ -partite graph is called a *co- $p$ -partite* graph. Observe that the vertex set of a co- $p$ -partite graph can be partitioned into at most  $p$  cliques. The complement of a bipartite graph is called *co-bipartite*.

### 3 Structure of $d$ -Dot Product Graphs

In this section, we describe some of the structure of  $d$ -dot product graphs, which we need in our algorithms later on. Fiduccia et al. [10, Theorem 20] proved that 1-dot product graphs have at most two nontrivial components, each of which are threshold graphs. We show that  $d$ -dot product graphs, and in particular 2-dot product graphs, exhibit similar interesting structural properties.

From now we assume that  $d \geq 2$ . The reason for doing this is that our polynomial-time results on INDEPENDENT SET and CLIQUE in Section 4 for the case  $d = 2$  readily carry over to the case  $d = 1$ : we can represent a  $(d - 1)$ -dot product graph as a  $d$ -dot product graph for all  $d \geq 2$  by adding a zero entry to all vectors of any of its  $(d - 1)$ -dot product representations.

We call a  $d$ -dot product representation of a graph *clean* if it contains no two vectors  $\mathbf{a}^u$  and  $\mathbf{a}^v$  with  $\mathbf{a}^u = \gamma \mathbf{a}^v$  for some  $\gamma \geq 0$ .

**Lemma 1.** (★)<sup>3</sup> *Given a  $d$ -dot product graph  $G$  without isolated vertices and a  $d$ -dot product representation of  $G$ , we can compute a clean  $d$ -dot product representation of  $G$  in polynomial time.*

<sup>3</sup> Proofs marked with a star have been omitted due to page restrictions.

Throughout the remainder of this section, we assume that we are given a  $d$ -dot product graph  $G = (V, E)$  for some  $d \geq 2$  together with a  $d$ -dot product representation with vectors  $\{\mathbf{a}^u \mid u \in V\}$  and threshold  $t$ . For solving INDEPENDENT SET and CLIQUE, we can preprocess  $G$  by removing any isolated vertices. Hence, by Lemma 1, we may assume without loss of generality that the given representation is clean.

We will use the notation  $\theta_{uv}$  for the angle between  $\mathbf{a}^u$  and  $\mathbf{a}^v$ , which is the smaller of the two angles between  $\mathbf{a}^u$  and  $\mathbf{a}^v$  in the plane defined by  $\mathbf{a}^u$  and  $\mathbf{a}^v$ . We assume some fixed direction of rotation so  $\theta_{uv} = -\theta_{vu}$ .

We say that a vertex  $u$  is *short* if  $\|\mathbf{a}^u\| \leq \sqrt{t}$ ; otherwise, it is *long*. Note that we can decide whether  $u$  is short in polynomial time by checking whether  $\|\mathbf{a}^u\|^2 \leq t$ . We first provide two lemmas about short vertices.

**Lemma 2.** (★) *Let  $v$  be a short vertex. Then  $G[N(v)]$  is co- $2^{d-1}$ -partite.*

The lemma shows in particular that  $G[N(v)]$  is co-bipartite if  $d = 2$ .

**Lemma 3.** (★) *The set of short vertices is an independent set.*

We say that a vertex  $v$  is *between* vertices  $u$  and  $w$  if  $\mathbf{a}^v$  can be written as a nonnegative linear combination of  $\mathbf{a}^u$  and  $\mathbf{a}^w$ . In other words,  $v$  is between  $u$  and  $w$  if  $\mathbf{a}^v$  lies in the plane defined by  $\mathbf{a}^u$  and  $\mathbf{a}^w$  and  $\mathbf{a}^v$  lies within the smaller of the two angles defined by  $\mathbf{a}^u$  and  $\mathbf{a}^w$  in this plane.

We now present two lemmas about the neighbourhoods of vertices.

**Lemma 4.** (★) *Let  $L = \{u \in V \mid \|\mathbf{a}^u\| > \sqrt{t}\}$ . If  $d = 2$ , then  $G[N(v) \cap L]$  is a co-comparability graph for all  $v \in V$ .*

**Lemma 5.** (★) *Let  $u, v, w \in V$  be such that  $v$  is between  $u$  and  $w$ . If  $u$  is adjacent to  $w$  and  $\|\mathbf{a}^v\| \geq \|\mathbf{a}^w\|$ , then  $u$  is adjacent to  $v$ .*

We also require a result that is implied by Lemma 28 of Fiduccia et al. [10].

**Lemma 6.** *Suppose  $d = 2$ . Let  $u, v$ , and  $w$  be vertices such that  $v$  is between  $u$  and  $w$ . If  $u$  is adjacent to  $w$ , and  $v$  is adjacent to neither  $u$  nor  $w$ , then  $v$  is short.*

## 4 Diversity and Clustering in Social Networks

In this section, we consider the complexity of computing our two measures of diversity and clustering in social networks, i.e. INDEPENDENT SET and CLIQUE, respectively, on a dot product graph model of the network. We first prove that INDEPENDENT SET is polynomial-time solvable if  $d \leq 2$  and NP-complete if  $d \geq 3$ . We then prove that CLIQUE is polynomial-time solvable if  $d \leq 2$ .

As before, throughout we have a  $d$ -dot product graph  $G = (V, E)$  and a clean  $d$ -dot product representation with vectors  $\{\mathbf{a}^u \mid u \in V\}$  and threshold  $t$ .

We first consider INDEPENDENT SET in the case  $d \leq 2$ . Recall that we may assume without loss of generality that  $d = 2$ . Armed with the structural results of the previous section, we can prove the following theorem.

**Theorem 1.** INDEPENDENT SET is solvable in  $O(n^3)$  time on 2-dot product graphs on  $n$  vertices.

*Proof.* Let  $G$  be a 2-dot product graph. We describe how to find a maximum size independent set of  $G$ . In fact, we will describe how to find, for each long vertex  $u$  of  $G$ , the maximum size independent set of  $G$  that contains  $u$ . This is sufficient as the maximum size set of  $G$  is either the largest of these sets, or the set of all short vertices which is also independent by Lemma 3; we use this latter fact repeatedly in this proof. So let  $u$  be a fixed long vertex of  $G$ . Let  $G_u$  be the graph obtained by removing all vertices that neighbour  $u$  and their incident edges. If we can find the maximum size independent set of  $G_u$ , we will have found the maximum size independent set of  $G$  that contains  $u$ .

We define a total (or linear) ordering  $\prec$  of the vertices of  $G_u$  by ordering the vertices by increasing angle of their vector representation from  $\mathbf{a}^u$ . Using the square of the cosine formula,  $\prec$  can be computed in quadratic time using just dot-products. We wish to relate this ordering to betweenness. Suppose that two vertices  $v$  and  $w$  are adjacent in  $G_u$  and that  $\theta_{vw}$  is positive. Any vertex between  $v$  and  $w$  is, by Lemma 6, either short or adjacent to one of them, and we know that  $u$  is a long vertex with no neighbours. So if  $x$  is between  $v$  and  $w$ , we have  $v \prec x \prec w$ . The converse is clearly true, giving us:

*Claim 1:* Let  $v, w, x$  be vertices in  $G_u$  where  $v$  and  $w$  are adjacent. Then  $x$  is between  $v$  and  $w$  and  $\theta_{vw}$  is positive if and only if  $v \prec x \prec w$ .

For a long vertex  $v$  in  $G_u$ , let  $J(v)$  be a largest independent set containing  $v$  in the subgraph of  $G_u$  that contains all vertices up to  $v$  in the ordering  $\prec$ , and let  $j(v) = |J(v)|$ . For a pair of long vertices  $v$  and  $w$  in  $G_u$  with  $w \prec v$ , let  $S(w, v)$  be the set of vertices  $x$  such that  $x$  is short,  $w \prec x \prec v$  and  $x$  is not adjacent to either  $v$  or  $w$ . Let  $s(w, v) = |S(w, v)|$ .

*Claim 2:* For each pair of non-adjacent long vertices  $v$  and  $w$  with  $w \prec v$  in  $G_u$ ,  $j(v) \geq j(w) + s(w, v) + 1$ .

*Proof.* Note that the claim will follow if we can show that  $J(w) \cup S(w, v) \cup \{v\}$  is an independent set. All we need to show is that no vertex in  $S(w, v) \cup \{v\}$  is adjacent to a vertex in  $J(w)$ .

Suppose that  $v$  is adjacent to a vertex  $x$  in  $J(w)$ . We know  $v$  and  $w$  are not adjacent so  $x \neq w$  and  $x \prec w \prec v$ . Hence,  $w$  is between  $x$  and  $v$  (by Claim 1), and the adjacency of  $x$  and  $v$  implies, by Lemma 6, that  $w$  is short; a contradiction.

If a vertex  $y \in S(w, v)$  is adjacent to any vertex  $x$  in  $J(w)$ , then  $x \neq w$  by the definition of  $S(w, v)$ . But  $x$  is adjacent to  $w$  using Lemma 5 and noting that  $w$  is long,  $y$  is short and  $w$  is between  $x$  and  $y$ . This contradiction proves Claim 2.

*Claim 3:* For each long vertex  $v \neq u$  in  $G_u$ ,  $j(v)$  is the maximum, over all long vertices  $w$  with  $w \prec v$  and  $v$  and  $w$  non-adjacent, of  $j(w) + s(w, v) + 1$ .

*Proof.* Note that the set of long vertices that precede  $v$  includes the isolated vertex  $u$  so the maximum is well-defined, and the previous claim tells us that

$j(v)$  is no less than this maximum. We must show that it is no larger. Let  $w$  be the long vertex that is last in the ordering amongst all long vertices in  $J(v) \setminus \{v\}$  (as  $J(v)$  contains  $u$  we can always find such a vertex). The subset of  $J(v)$  containing only  $w$  and preceding vertices is independent and contains at most  $j(w)$  vertices. The only other vertices in  $J(v)$  are short vertices between  $w$  and  $v$  and  $v$  itself. Thus  $j(v) \leq j(w) + s(w, v) + 1$ , and Claim 3 is proved.

Note that  $j$  can easily be computed since  $j(u) = 1$ , and Claim 3 tells us that if we consider the vertices in order we can find the remaining values.

For each long vertex  $v$  in  $G_u$ , let  $S^+(v)$  contain each vertex  $w$  such that  $w$  is short,  $v \prec w$  and  $v$  is not adjacent to  $w$ . Let  $s^+(v) = |S^+(v)|$ . Let  $m$  be the maximum, over all long vertices  $v$  in  $G_u$ , of  $j(v) + s^+(v)$ .

*Claim 4:* Let  $J$  be a maximum size independent set in  $G_u$ . Then  $|J| = m$ .

*Proof.* Let  $v$  be a long vertex in  $G_u$ . We shall show that  $J(v) \cup S^+(v)$  is an independent set. Let  $w$  be a vertex in  $S^+(v)$  and suppose that  $x$  is a vertex in  $J(v)$  adjacent to  $w$ . By the definition of  $S^+(v)$ , we have  $x \neq v$ , so  $x \prec v \prec w$ . By Claim 1,  $v$  is between  $x$  and  $w$  and, by Lemma 6,  $v$  is either short or adjacent to  $x$  or  $w$ . This contradiction shows that  $J(v) \cup S^+(v)$  is an independent set. So  $|J| \geq j(v) + s^+(v)$  for all long vertices  $v$  and hence  $|J|$  is at least  $m$ .

Now let  $z$  be the long vertex in  $J$  that is latest in the ordering. Let  $J_1$  be the subset of  $J$  containing  $z$  and preceding vertices. Hence,  $|J_1| \leq j(z)$ . The vertices of  $J \setminus J_1$  are short vertices later than  $z$  in the ordering, so there are at most  $s^+(z)$  of them. Thus  $|J| \leq j(z) + s^+(z) \leq m$ , and Claim 4 is proved.

We omit the details but it is straightforward to show that  $j$  and  $s^+$ , and so also  $m$ , can be computed in  $O(n^2)$  time. The corresponding sets of vertices, and thus a maximum size independent set of  $G_u$ , can also be found. By repeating for each  $u$ , a maximum size independent set of  $G$  is found in time  $O(n^3)$ .  $\square$

We contrast this positive result by the following result.

**Theorem 2.** (★) *For any  $d \geq 3$ , INDEPENDENT SET is NP-complete on  $d$ -dot product graphs<sup>4</sup>.*

The structural results of the previous section provide enough structure to solve CLIQUE in polynomial time on 2-dot product graphs.

**Theorem 3.** (★) *CLIQUE is solvable in  $O(n^4)$  time on 2-dot product graphs on  $n$  vertices, even if no 2-dot product representation is given.*

## 5 Structure and Complexity for Variants of the Model

In this section, we consider two variants of the dot product graph model, which model that two individuals are connected if and only if their preferences are not

<sup>4</sup> Here the problem input consists of the graph, but not (necessarily) a representation.

opposite, or are neither opposite nor orthogonal. In the introduction, we defined the  $d^0$ -dot product graph and the  $d^+$ -dot product graph model for these cases. Recall that if  $\{\mathbf{a}^u \mid u \in V\}$  is a representation of  $G = (V, E)$ , then

- $(u, v) \in E$  if and only if  $\mathbf{a}^u \cdot \mathbf{a}^v \geq 0$  when  $G$  is a  $d^0$ -dot product graph, and
- $(u, v) \in E$  if and only if  $\mathbf{a}^u \cdot \mathbf{a}^v > 0$  when  $G$  is a  $d^+$ -dot product graph.

We study the complexity of computing the diversity and clustering measures on these models, that is, of INDEPENDENT SET and CLIQUE, on  $d^0$ -dot product graphs and  $d^+$ -dot product graphs.

Note that vertices of length 0 are adjacent to all other vertices in a  $d^0$ -dot product graph and are isolated in a  $d^+$ -dot product graph, and so do not, in either case, influence INDEPENDENT SET or CLIQUE. Hence, without loss of generality all vectors in this section have non-zero length.

First, we describe the structure of independent sets in  $d^0$ -dot product graphs. The following lemma is equivalent to Lemma 18 of Fiduccia et al. [10].

**Lemma 7.** *For all  $d \geq 1$ , every independent set in a  $d^0$ -dot product graph has size at most  $d + 1$ .*

Independent sets in  $d^+$ -dot product graphs have a different structure.

**Lemma 8. (★)** *For all  $d \geq 1$ , every independent set in a  $d^+$ -dot product graph has size at most  $2d$ .*

The proofs of Lemmas 7 and 8 can be turned into constructions to show that the given bounds are tight. The lemmas show that  $d^0$ -dot product graphs and  $d^+$ -dot product graphs have different structure, which is also different from the structure of  $d$ -dot product graphs. Moreover, using exhaustive enumeration, the two lemmas immediately imply the following.

**Theorem 4.** *For all  $d \geq 1$ , INDEPENDENT SET is solvable in  $O(n^{d+1})$  time on  $d^0$ -dot product graphs and in  $O(n^{2d})$  time on  $d^+$ -dot product graphs on  $n$  vertices, even if no representation is given.*

We now consider CLIQUE on  $d^0$ -dot product and  $d^+$ -dot product graphs. For  $d = 2$ , it suffices to observe that a set of vertices forms a clique if and only if their corresponding vectors lie in the nonnegative quadrant (after an appropriate rotation). However, this structural observation does not generalize to higher dimensions, as is evident from the counterexamples by Gray and Wilson [12] for  $d = 3$  and  $d \geq 5$ . Instead, we follow a different approach, which leads to a polynomial-time algorithm for all  $d \leq 3$ .

For any hyperplane  $h$  with normal  $\mathbf{n}$ , let  $h^+$  be the half-space  $\{p \mid p \cdot \mathbf{n} \geq 0\}$  and let  $h^-$  be the half-space  $\{p \mid p \cdot \mathbf{n} \leq 0\}$ . Note that any two vectors  $\mathbf{a}, \mathbf{b}$  induce a hyperplane with normal  $\mathbf{a} \times \mathbf{b}$ , where  $\times$  is the cross product operation. We refer to the monograph by Barvinok [4] for any undefined terminology on cones.

**Theorem 5.** *For all  $d \leq 3$ , CLIQUE can be solved in  $O(n^{4.5})$  time on  $d^0$ -dot product graphs and  $d^+$ -dot product graphs on  $n$  vertices.*

*Proof.* We assume that  $d = 3$  (fewer dimensions are a special case). Let  $G = (V, E)$  be a  $3^0$ -dot product graph or a  $3^+$ -dot product graph with representation  $\{\mathbf{a}^v \mid v \in V\}$ . We first give a structural result, where we essentially show that any clique  $C$  of  $G$  induces a basis such that the vectors of  $C$  lie in two octants with respect to this basis. Then, we give an algorithm that finds this basis for a maximum clique by guessing limited information about the clique, and uses the basis to obtain a maximum clique of  $G$ .

We start with the structural result. Let  $C$  be any clique of  $G$ . Let  $\mathcal{K}$  denote the conic hull of  $\mathbf{a}^v$  for all vertices  $v \in C$ , that is,  $\mathcal{K} = \{\sum_{v \in C} \lambda_v \mathbf{a}^v \mid \lambda_v \geq 0\}$ . We call  $\mathcal{K}$  the *cone corresponding to  $C$* . The structural result considers the case that  $\mathcal{K}$  is not a ray (a *ray* is the conic hull of a single vector). Since  $\mathcal{K}$  is generated by a finite set, its extreme rays are vectors that correspond to vertices of  $C$ . Let  $u$  be any vertex such that  $\mathbf{a}^u$  spans an extreme ray of  $\mathcal{K}$ , and let  $h_u$  denote the hyperplane with normal  $\mathbf{a}^u$ . Because  $\mathcal{K}$  is the conic hull of vectors corresponding to a clique,  $\mathbf{p} \cdot \mathbf{a}^u \geq 0$  for any  $\mathbf{p} \in \mathcal{K}$  (this is true both when  $G$  is a  $3^0$ -dot product graph or a  $3^+$ -dot product graph). Hence,  $\mathcal{K} \subseteq h_u^+$ .

Let  $w$  be any vertex such that  $\mathbf{a}^w$  spans an extreme ray of  $\mathcal{K}$  that is not spanned by  $u$  and such that the hyperplane  $h_{uw}$  induced by  $\mathbf{a}^u$  and  $\mathbf{a}^w$  contains a facet of  $\mathcal{K}$ . Since  $h_{uw}$  contains a facet of  $\mathcal{K}$ , either  $\mathcal{K} \subseteq h_{uw}^+$  or  $\mathcal{K} \subseteq h_{uw}^-$ . Assume without loss of generality that  $\mathcal{K} \subseteq h_{uw}^+$ , and let  $\mathbf{t}$  denote the normal of  $h_{uw}$  that lies in  $h_{uw}^+$ . Finally, let  $\mathbf{w}'$  denote the projection of  $\mathbf{a}^w$  onto  $h_u$ . By definition,  $\mathbf{t}$ ,  $\mathbf{a}^u$ ,  $\mathbf{w}'$  are pairwise orthogonal. Moreover, as  $\mathcal{K} \subseteq h_u^+ \cap h_{uw}^+$  and  $h_u^+ \cap h_{uw}^+$  is the union of two octants in the basis induced by  $\mathbf{t}, \mathbf{a}^u, \mathbf{w}'$ , we find that  $\mathcal{K}$  is a subset of two octants in the basis induced by  $\mathbf{t}, \mathbf{a}^u, \mathbf{w}'$ .

We use the structural result in an algorithm that consists of two phases.

In the first phase of the algorithm, we ensure that we find a maximum clique if the cone corresponding to some maximum clique is a ray. Therefore, we iterate over all  $v \in V(G)$  and find the set  $X$  of vertices  $u$  for which  $\mathbf{a}^u$  spans the same ray as  $\mathbf{a}^v$ . The set  $X$  is a clique irrespective of whether  $G$  is a  $3^0$ -dot product graph or a  $3^+$ -dot product graph. We keep a maximum clique found over all choices of  $v$ .

In the second phase of the algorithm, we ensure that we find a maximum clique if the cone corresponding to some maximum clique is not a ray. Iterate over all  $n^2$  ordered pairs  $(u, w)$  of the vertices of  $G$  such that  $\mathbf{a}^u$  and  $\mathbf{a}^w$  do not span the same ray. Define  $h_u$  as the plane with normal  $\mathbf{a}^u$ , and define  $h_{uw}$  as the plane induced by  $\mathbf{a}^u$  and  $\mathbf{a}^w$ . Consider  $h_u^+ \cap h_{uw}^+$  (we also consider  $h_u^+ \cap h_{uw}^-$  in a similar way). Let  $\mathbf{t}$  denote the normal of  $h_{uw}$  that lies in  $h_{uw}^+$  and let  $\mathbf{w}'$  denote the projection of  $\mathbf{a}^w$  onto  $h_u$ . Note that  $h_u^+ \cap h_{uw}^+$  is the union of two octants in the basis induced by  $\mathbf{t}, \mathbf{a}^u, \mathbf{w}'$ . As any octant induces a clique,  $h_u^+ \cap h_{uw}^+$  induces a co-bipartite graph  $H$ . We can find  $H$  in linear time as the graph induced by the vertices whose corresponding vectors have positive or strictly positive dot product with both  $\mathbf{a}^u$  and  $\mathbf{t}$ . Since  $H$  is co-bipartite, we can find a maximum clique of  $H$  in  $O(n^{2.5})$  time, as it reduces to finding a maximum matching in a bipartite graph, which takes  $O(n^{2.5})$  time [15]. We then keep a maximum clique over all choices of  $u, w$ . The output of the algorithm is a largest of the two cliques kept in the first and second phase.

The algorithm runs in  $O(n^{4.5})$  time, as claimed. To see correctness, let  $C$  be a maximum clique. If the cone corresponding to  $C$  is a ray, then the algorithm considers  $C$  in the first phase. If the cone corresponding to  $C$  is not a ray, then by our structural result there will be a choice of  $u, w$  for which  $u, w \in C$  and  $h_{uw}$  contains a facet of  $\mathcal{K}$ , where  $\mathcal{K}$  is the cone corresponding to  $C$ .  $\square$

## 6 Conclusions

This paper provided the first study of algorithms that measure diversity and clustering in social networks that are modeled as dot product graphs. The diversity and clustering measures considered correspond to INDEPENDENT SET and CLIQUE on dot product graphs.

Our exploration of the complexity of CLIQUE on  $d$ -dot product graphs leaves further open problems. The current approach for  $d = 2$  does not seem to extend to  $d$ -dot product graphs for  $d \geq 3$ , as our structural results (Lemma 2 for example) seem to indicate that we need to solve clique on co- $p$ -partite graphs for  $p \geq 3$ . However, this problem is NP-complete, as INDEPENDENT SET is NP-complete on 2-subdivisions of planar graphs [11]. Hence, further structural insight into  $d$ -dot product graphs is needed to resolve the complexity of CLIQUE on these graphs.

We observe that our polynomial-time algorithms for INDEPENDENT SET and CLIQUE on 2-dot product graphs generalize well-known polynomial-time algorithms for these problems on interval graphs, because interval graphs have a 2-dot product representation [10, Theorem 21]. At the same time, we are unaware of any nontrivial superclasses of 2-dot product graphs, in particular for which INDEPENDENT SET and CLIQUE are polynomial-time solvable. Finally, we note that the dot product graph model of social networks might be able to capture more problems for social networks as graph optimization problems.

## References

1. L.A. Adamic, E. Adar, Friends and neighbors on the Web, *Social Networks* 25 (2003) 211–230.
2. S. Arora, D. Steurer, A. Wigderson, Towards a Study of Low-Complexity Graphs, *Proc. ICALP 2009, LNCS 5555* (2009) 119–131.
3. A.-L. Barabási and R. Albert, Emergence of Scaling in Random Networks, *Science* 286 (1999) 509–512.
4. A. Barvinok, *A Course in Convexity*, American Mathematical Society, 2003.
5. R.L. Breiger, The Duality of Persons and Groups, *Social Forces* 53 (1974) 181–190.
6. K. Bhawalkar, J.M. Kleinberg, K. Lewi, T. Roughgarden and A. Sharma, Preventing Unraveling in Social Networks: The Anchored k-Core Problem, *Proc. ICALP 2012, LNCS 7392* (2012) 440–451.
7. G. Caldarelli, A. Capocci, P. de Los Rios, M.A. Muñoz, Scale-Free Networks from Varying Vertex Intrinsic Fitness, *Phys. Rev. Lett.* 89 (2002), 258702.
8. R. Diestel, *Graph Theory*, Springer-Verlag, 2005.
9. R.G. Downey and M.R. Fellows, Fixed-parameter tractability and completeness II: On completeness for W[1], *Theoretical Computer Science* 141 (1995) 109–131.

10. C.M. Fiduccia, E.R. Scheinerman, A. Trenk and J.S. Zito, Dot product representations of graphs, *Discrete Mathematics* 181 (1998) 113–138.
11. M.R. Garey, D.S. Johnson and L. Stockmeyer, Some simplified NP-complete graph problems, *Theor. Comp. Sci.* 1 (1976) 237–267.
12. L.J. Gray and D.G. Wilson, Nonnegative factorization of positive semidefinite non-negative matrices, *Linear Algebra and its Applications* 31 (1980) 119–127.
13. J. Håstad, Clique is hard to approximate within  $n^{1-\epsilon}$ , *Acta Mathematica* 182 (1999) 105–142.
14. P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent Space Approaches to Social Network Analysis, *J. Am. Stat. Assoc.* 97 (2002) 1090–1098.
15. J.E. Hopcroft, R.M. Karp, An  $n^{5/2}$  Algorithm for Maximum Matchings in Bipartite Graphs, *SIAM J. Comput.* 2 (1973) 225–231.
16. R.J. Kang, L. Lovász, T. Müller and E.R. Scheinerman, Dot product representations of planar graphs. *Electr. J. Comb.* 18 (2011).
17. R.J. Kang and T. Müller, Sphere and dot product representations of graphs, *Discrete and Computational Geometry* 47 (2012) 548–568.
18. R.M. Karp, “Reducibility among Combinatorial Problems”, *Complexity of Computer Computations*, Plenum Press (1972) 85–103.
19. M. Kim, J. Leskovec, Multiplicative Attribute Graph Model of Real-World Networks, *Proc. WAW 2010, LNCS 6516* (2010) 62–73.
20. A. Kotlov, L. Lovász and S. Vempala, The Colin de Verdière number and sphere representations of a graph, *Combinatorica* 17 (1997) 483–521.
21. D. Liben-Nowell, J. Kleinberg, The Link Prediction Problem for Social Networks, *Proc. CIKM 2003* (2003) 556–559.
22. M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a Feather: Homophily in Social Networks, *Annual Rev. Sociology* 27 (2001) 415–444.
23. M.E.J. Newman, The Structure and Function of Complex Networks, *SIAM Review* 45 (2003) 167–256.
24. C.M.L. Nickel, Random Dot Product Graphs: A Model for Social Networks, PhD dissertation, Johns Hopkins University, 2007.
25. C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent Semantic Indexing: A Probabilistic Analysis, *J. Comput. Syst. Sci.* 61 (2000) 217–235.
26. D.J. de S. Price, A general theory of bibliometric and other cumulative advantage processes, *J. American Society for Information Science* 27 (1976) 292–306.
27. J. Reiterman, V. Rödl and E. Šinǎjová, Embeddings of graphs in Euclidean spaces, *Discrete and Computational Geometry* 4 (1989) 349–364.
28. J. Reiterman, V. Rödl and E. Šinǎjová, Geometrical embeddings of graphs. *Discrete Mathematics* 74 (1989) 291–319.
29. J. Reiterman, V. Rödl and E. Šinǎjová, On embedding of graphs into Euclidean spaces of small dimension, *J. Combin. Theory B* 56 (1992) 1–8.
30. E.R. Scheinerman, K. Tucker, Modeling graphs using dot product representations, *Computational Statistics* 25 (2010) 1–16.
31. H.A. Simon, On a class of skew distribution functions, *Biom.* 42 (1955) 425–440.
32. T.A.B. Snijders, Statistical Models for Social Networks, *Annual Rev. Sociology* 37 (2011) 131–153.
33. D.J. Watts, P.S. Dodds, M.E.J. Newman, Identity and Search in Social Networks, *Science* 296 (2002) 1302–1305.
34. S.J. Young and E.R. Scheinerman, Random dot product graph models for social networks, *Proc. WAW 2007, LNCS 4863* (2007) 138–149.
35. S.J. Young and E.R. Scheinerman, Directed random dot product graphs, *Internet Mathematics* 5(2008) 91–111.