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

Matthew Johnson[1], Daniël Paulusma[1], and Erik Jan van Leeuwen[2]

[1] School of Engineering and Computer Science, Durham University, England
{matthew.johnson2,daniel.paulusma}@durham.ac.uk
[2] 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.

Keywords. social network; d-dot product graph; independent set; clique.

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 e.g. the surveys by Newman [28] and Snijders [37]). Many of these studies verify that a particular model has properties

* This paper was supported by EPSRC (EP/G043434/1), and an extended abstract of it will appear in the Proceedings of ISAAC 2013.

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 [26]). For example, the models of Simon [36], Price [31], 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 e.g. the survey by McPherson et al. [27]). 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 [38]. 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 e.g. [1, 17, 22, 38]). 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

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

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 graph as a model for social networks was recently formalized by Nickel, Scheinerman, Tucker, and Young [35, 39, 40, 29]. Their studies were motivated by earlier work of Papadimitriou et al. [30] and Caldarelli et al. [7]. Moreover, the dot product similarity measure is similar to the cosine measure, which was studied in information retrieval and social networks before (see e.g. [8, 9]). We note, however, that dot product graphs have a much longer tradition, both in sociology (see e.g. Breiger [5]) and in graph theory. We briefly survey known graph-theoretic results. Reiterman et al. [32–34] and particularly Fiduccia et al. [12] proved several structural results. The work of Fiduccia et al. [12] implies that 1-dot product graphs can be recognized in polynomial time. However, Kang and Müller [20] 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) [23, 19], deriving for example a tight bound of 4 on the dot product dimension of a planar graph [19]. Fiduccia et al. [12] conjectured that any graph on n vertices has dot product dimension at most $\frac{n}{2}$; Li and Chang [25] recently confirmed this conjecture for a number of graph classes. 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 coefficient, 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 [21, 11, 16], 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 NP-complete, W[1]-complete, and very hard to approximate on general graphs [21, 11, 16], 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. [12] that the graph class where two vertices are adjacent if and only if $\mathbf{a}^u \cdot \mathbf{a}^v >$

Setting	Independent Set	Clique
d-DPG (≥ 1)	in P for $d \leq 2$	in P for $d \leq 2$
	NP-complete for $d \geq 3$	? for $d \geq 3$
d^0-DPG (≥ 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.

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.

Organization. In Section 3, we prove several structural results about d-dot product graphs. In Section 4, we consider the complexity of Independent Set and Clique on dot product graphs. In Section 5, we study the computational complexity of these problems on d^0- and d^+-dot product graphs.

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 [10].

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*.

The *length* of a cycle is its number of edges. The *girth* of a graph G is the length of a shortest induced cycle in G.

3 Structure of *d*-Dot Product Graphs

In this section, we describe some of the structure that can be found in d-dot product graphs and which we need in our algorithms later on. Fiduccia et al. [12, 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. *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.*

Proof. Let $G = (V, E)$ be a d-dot product graph, and let $\{\mathbf{a}^u \mid u \in V\}$ be a d-dot product representation of G. Let t be the threshold. Recall that for any vertex u we may assume without loss of generality that $\mathbf{a}^u \cdot \mathbf{a}^v \neq t$ for any other vertex v [12, Proposition 2] and so, given u, we can find $\epsilon > 0$ such that $|t - \mathbf{a}^u.\mathbf{a}^v| > \epsilon$ for all v. And if $\mathbf{e}$ is a vector such that $|\mathbf{e}| < \epsilon/|\mathbf{a}^{v^*}|$ (where v^* is the longest vertex in the graph), then, for all v, the simple observation that

$$(\mathbf{a}^u + \mathbf{e}).\mathbf{a}^v = \mathbf{a}^u.\mathbf{a}^v + \mathbf{e}.\mathbf{a}^v$$

implies that $(\mathbf{a}^u + \mathbf{e}).\mathbf{a}^v \geq t$ if and only if $\mathbf{a}^u.\mathbf{a}^v \geq t$. So we can replace $\mathbf{a}^u$ by $\mathbf{a}^u + \mathbf{e}$ to obtain another d-dot product representation.

Hence to find our clean representation, we consider each pair of vertices u, v in turn and if $\mathbf{a}^u = \gamma \mathbf{a}^v$ for some $\gamma \geq 0$ we just replace $\mathbf{a}^u$ by an appropriate $\mathbf{a}^u + \mathbf{e}$ (and if by some bad luck $\mathbf{e}$ is chosen such that $\mathbf{a}^u + \mathbf{e} = \gamma' \mathbf{a}^w$ for another vertex w, we note that, for example, $\mathbf{e}/2$ or $-\mathbf{e}$ can be used in place of $\mathbf{e}$). Without considering the computations needed in detail, it is clear that the representation can be obtained in polynomial time. □

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 θ_{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.*

Proof. We may assume that the representation is rotated such that $\mathbf{a}^v_1 = z$ for some $0 < z \leq \sqrt{t}$ and that $\mathbf{a}^v_i = 0$ for all $i = 2, \ldots, d$, i.e., that $\mathbf{a}^v$ is the d-dimensional unit vector scaled by some $z > 0$. Observe that $u \in N(v)$ if and only if $\mathbf{a}^u_1 \geq t/z \geq \sqrt{t}$. Associate with each vertex $u \in N(v)$ a $(d-1)$-dimensional sign vector $\boldsymbol{s}^u$, where $\boldsymbol{s}^u_i = 1$ if $\mathbf{a}^u_{i+1} \geq 0$ and $\boldsymbol{s}^u_i = -1$ otherwise for $i = 1, \ldots, d-1$. Observe that the sign-vectors naturally partition the vertices of $N(v)$ into 2^{d-1} equivalence classes. Moreover, any two vertices u, w in an equivalence class are adjacent, because $\mathbf{a}^u \cdot \mathbf{a}^w = \sum_{i=1}^{d} \mathbf{a}^u_i \mathbf{a}^w_i \geq t + \sum_{i=2}^{d} \mathbf{a}^u_i \mathbf{a}^w_i \geq t$, as $\mathbf{a}^u_i \geq 0$ if and only if $\mathbf{a}^w_i \geq 0$ for any $i = 2, \ldots, d$. Therefore, each equivalence class induces a clique, and thus $G[N(v)]$ is co-2^{d-1}-partite. □

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

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

Proof. For any two vertices v and w, we have $\boldsymbol{a}^v \cdot \boldsymbol{a}^w = \|\boldsymbol{a}^v\| \, \|\boldsymbol{a}^w\| \cos\theta_{vw}$. Assume that v and w are short. As we assume that our d-dot product representation of G is clean, $\cos\theta_{vw} < 1$ which, combined with the definition of short, implies that $\boldsymbol{a}^v \cdot \boldsymbol{a}^w < t$, and hence the vertices are not adjacent. □

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

We require a result that in the case that $t = 1$ is implied by Lemma 28 of Fiduccia et al. [12]. The generalization to all t can easily be obtained by copying their proof, so here we will state it without proof.

Lemma 4. *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.*

We now present two lemmas about the neighbourhoods of vertices.

Lemma 5. *Let $L = \{u \in V \mid \|\boldsymbol{a}^u\| > \sqrt{t}\}$. If $d = 2$, then $G[N(v) \cap L]$ is a co-comparability graph for all $v \in V$.*

Proof. We may assume that the representation is rotated such that $\boldsymbol{a}^v = (z, 0)$ for some $z > 0$. Then we have $u \in N(v)$ only if $\boldsymbol{a}^u_1 \geq 0$. Number the vertices of $N(v) \cap L$ by increasing angle with respect to the vector $(0, 1)$. Let u_i denote the i-th vertex in the order. Consider some $i < j < k$ such that $(u_i, u_k) \in E$. Note that $\boldsymbol{a}^{u_j}$ is between $\boldsymbol{a}^{u_i}$ and $\boldsymbol{a}^{u_k}$. Since $u_j \in L$, it follows from Lemma 4 that one of $(u_i, u_j), (u_k, u_j) \in E$. The existence of such an ordering implies that $G[N(v) \cap L]$ is a co-comparability graph, due to Kratsch and Stewart [24]. □

Lemma 6. *Let $u, v, w \in V$ be such that v is between u and w. If u is adjacent to w and $\|\boldsymbol{a}^v\| \geq \|\boldsymbol{a}^w\|$, then u is adjacent to v.*

Proof. Notice that $\cos\theta_{uv} > \cos\theta_{uw}$. Hence,

$$\begin{aligned}\boldsymbol{a}^u \cdot \boldsymbol{a}^v &= \|\boldsymbol{a}^u\| \, \|\boldsymbol{a}^v\| \cos\theta_{uv} \\ &> \|\boldsymbol{a}^u\| \, \|\boldsymbol{a}^w\| \cos\theta_{uw} \\ &= \boldsymbol{a}^u \cdot \boldsymbol{a}^w \\ &\geq t\end{aligned}$$

and so u and v are adjacent. □

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 $\boldsymbol{a}^u$. 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 θ_{vw} is positive. Any vertex between v and w is, by Lemma 4, 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 θ_{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 4, 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 6 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 4, 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 $O(n^3)$. □

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* [3].

Proof. The Independent Set problem is NP-complete on planar graphs [14]. If Independent Set is NP-complete on a particular graph class $\mathcal{G}$, it remains NP-complete on the graph class obtained by 2-subdividing each edge of each graph of $\mathcal{G}$; recall that an s-subdivision of an edge is the operation in which the edge is replaced by an $(s+1)$-edge path (implying that s new vertices are created). Hence, Independent Set is NP-complete on 2-subdivisions of planar graphs. Note that such graphs are planar and have girth at least 9. Kang et al. [19] observed that planar graphs of girth at least 5 are 3-dot product graphs. So Independent Set is NP-complete on 3-dot product graphs which are a subclass of d-dot product graphs for $d > 3$. □

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.*

Proof. Call a vertex v *weak* if $N(v)$ is co-bipartite. Note that we can determine whether a vertex v is weak in quadratic time. We first find a largest clique that contains a weak vertex. Observe that Clique can be solved in $O(n^{2.5})$ time on co-bipartite graphs, since it reduces to finding a maximum matching of a bipartite graph, which can be solved in $O(n^{2.5})$ time [18]. Thus we can find the largest clique containing a weak vertex in $O(n^{3.5})$ time.

We then find a largest clique that only contains vertices that are not weak. Let L be the set of vertices that are not weak. By Lemma 2, each vertex of L must be long in some 2-dot product representation of the graph. Hence, for any $v \in L$, $G[N(v) \cap L]$ is a co-comparability graph by Lemma 5. We now observe that Clique on co-comparability graphs is Independent Set on comparability graphs. The latter problem can be reduced to a maximum-flow computation [13], which takes $O(n^3)$ time. Hence, we can find this largest clique in $O(n^4)$ time.

[3] Here the problem input consists of the graph, but not (necessarily) a representation.

Finally, we return the largest of the two cliques that we found. Note that we only used a 2-dot product representation in the analysis, and the algorithm itself also works 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 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. [12].

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 $2d$.*

Proof. We apply induction on d. Let $G = (V, E)$ be a d^+-dot product graph with representation $\{\mathbf{a}^u \mid u \in V\}$. The lemma is readily seen to hold for $d = 1$.

Let $d \geq 2$ and suppose that the claim holds for dimension $d - 1$. Let $I = \{u_1, \ldots, u_p\}$ for some $p \geq 1$ be an independent set of G. Without loss of generality, $\mathbf{a}^{u_1} = (1, 0, \ldots, 0)$. Consider the $(d-1)^+$-dot product graph $G' = (V', E')$ obtained from G by removing all first coordinates from the vectors $\mathbf{a}$ and then removing vectors of zero length. We claim that $I \cap V'$ is an independent set in G'. To see this, let $v, w \in I \cap V'$. Since v, w are independent of u_1, $\mathbf{a}_1^v, \mathbf{a}_1^w \leq 0$, and thus $\mathbf{a}_1^v \mathbf{a}_1^w \geq 0$. As $v, w \in I$, $\mathbf{a}^v \cdot \mathbf{a}^w \leq 0$ and thus $\sum_{i=2}^{d} \mathbf{a}_i^v \mathbf{a}_i^w \leq -(\mathbf{a}_1^v \mathbf{a}_1^w) \leq 0$. Hence, $\sum_{i=2}^{d} \mathbf{a}_i^v \mathbf{a}_i^w \leq 0$. Therefore, $I \cap V'$ is indeed an independent set.

By induction, we find that $|I \cap V'| \leq 2d - 2$. Notice that the only vertices w that are in G but not in G' are those for which $\mathbf{a}_i^w = 0$ for $i = 2, \ldots, d$. Suppose that $I \setminus \{u_1\}$ contains two such vertices, say v, w. They must satisfy $\mathbf{a}_1^v, \mathbf{a}_1^w < 0$ in order to be independent from u_1. It follows that $\mathbf{a}^v \cdot \mathbf{a}^w > 0$ and thus that v and w are adjacent, a contradiction. Therefore $I \setminus \{u_1\}$ can contain at most one vertex that is in G but not in G'. Hence $|I| \leq |I \cap V'| + 2 \leq 2d$. □

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 evident from the counterexamples by Gray and Wilson [15] for $d = 3$ and $d \geq 5$; see Appendix A for a counterexample for the case $d = 4$. 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 note that if a basis change is applied, then the resulting vectors are still a representation of the same kind (3^0-dot product or 3^+-dot product) for G. 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 use 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. 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 now turn the insight of the structural result into an algorithm. The algorithm 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_w. 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 $\mathbf{a}^v$ 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 [18]. 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.

Note that 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, and thus outputs a clique of size $|C|$. 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. For this choice of u, w, the algorithm finds a clique of size $|C|$. Hence, our algorithm outputs a maximum clique of G. □

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. We focussed on classical complexity only, and both approximability and parameterized complexity remain largely unexplored.

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 (e.g., Lemma 2) 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 [14]. 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 [12, 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. C. Cattuto, D. Benz, A. Hotho, G. Stumme, Semantic Grounding of Tag Relatedness in Social Bookmarking Systems, Proc. ISWC 2008, LNCS 5318 (2008) 615–631.
9. D.J. Crandall, D. Cosley, D.P. Huttenlocher, J.M. Kleinberg, S.Suri, Feedback effects between similarity and social influence in online communities, Proc. KDD 2008, ACM (2008) 160–168.
10. R. Diestel, Graph Theory, Springer-Verlag, 2005.
11. 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.
12. C.M. Fiduccia, E.R. Scheinerman, A. Trenk and J.S. Zito, Dot product representations of graphs, Discrete Mathematics 181 (1998) 113–138.
13. M.C. Golumbic, Algorithmic Graph Theory and Perfect Graphs, Acad. Press, 1988.
14. M.R. Garey, D.S. Johnson and L. Stockmeyer, Some simplified NP-complete graph problems, Theor. Comp. Sci. 1 (1976) 237–267.
15. L.J. Gray and D.G. Wilson, Nonnegative factorization of positive semidefinite nonnegative matrices, Linear Algebra and its Applications 31 (1980) 119–127.
16. J. Håstad, Clique is hard to approximate within $n^{1-\epsilon}$, Acta Mathematica 182 (1999) 105–142.

17. P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent Space Approaches to Social Network Analysis, J. Am. Stat. Assoc. 97 (2002) 1090–1098.
18. J.E. Hopcroft, R.M. Karp, An $n^{5/2}$ Algorithm for Maximum Matchings in Bipartite Graphs, SIAM J. Comput. 2 (1973) 225–231.
19. R.J. Kang, L. Lovász, T. Müller and E.R. Scheinerman, Dot product representations of planar graphs. Electr. J. Comb. 18 (2011).
20. R.J. Kang and T. Müller, Sphere and dot product representations of graphs, Discrete and Computational Geometry 47 (2012) 548–568.
21. R.M. Karp, Reducibility among Combinatorial Problems", Complexity of Computer Computations, Plenum Press (1972) 85–103.
22. M. Kim, J. Leskovec, Multiplicative Attribute Graph Model of Real-World Networks, Proc. WAW 2010, LNCS 6516 (2010) 62–73.
23. A. Kotlov, L. Lovász and S. Vempala, The Colin de Verdiére number and sphere representations of a graph, Combinatorica 17 (1997) 483–521.
24. D. Kratsch and L. Stewart, Domination on cocomparability graphs, SIAM J. Disc. Math. 6 (1993) 400–417.
25. B. Li and G.J. Chang, Dot product dimensions of graphs, Discrete Applied Mathematics, to appear.
26. D. Liben-Nowell, J. Kleinberg, The Link Prediction Problem for Social Networks, Proc. CIKM 2003 (2003) 556–559.
27. M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a Feather: Homophily in Social Networks, Annual Rev. Sociology 27 (2001) 415–444.
28. M.E.J. Newman, The Structure and Function of Complex Networks, SIAM Review 45 (2003) 167–256.
29. C.M.L. Nickel, Random Dot Product Graphs: A Model for Social Networks, PhD dissertation, Johns Hopkins University, 2007.
30. C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent Semantic Indexing: A Probabilistic Analysis, J. Comput. Syst. Sci. 61 (2000) 217–235.
31. 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.
32. J. Reiterman, V. Rödl and E. Šinăjová, Embeddings of graphs in Euclidean spaces, Discrete and Computational Geometery 4 (1989) 349–364.
33. J. Reiterman, V. Rödl and E. Šinăjová, Geometrical embeddings of graphs. Discrete Mathematics 74 (1989) 291–319.
34. J. Reiterman, V. Rödl and E. Šinăjová, On embedding of graphs into Euclidean spaces of small dimension, J. Combin. Theory B 56 (1992) 1–8.
35. E.R. Scheinerman, K. Tucker, Modeling graphs using dot product representations, Computational Statistics 25 (2010) 1–16.
36. H.A. Simon, On a class of skew distribution functions, Biom. 42 (1955) 425–440.
37. T.A.B. Snijders, Statistical Models for Social Networks, Annual Rev. Sociology 37 (2011) 131–153.
38. D.J. Watts, P.S. Dodds, M.E.J. Newman, Identity and Search in Social Networks, Science 296 (2002) 1302–1305.
39. S.J. Young and E.R. Scheinerman, Random dot product graph models for social networks, Proc. WAW 2007, LNCS 4863 (2007) 138–149.
40. S.J. Young and E.R. Scheinerman, Directed random dot product graphs, Internet Mathematics 5(2008) 91–111.

A Counterexample for the Case $d = 4$

Gray and Wilson [15] showed there exist sets of four vectors in $\mathbb{R}^3$ and sets of $\lfloor d/2 \rfloor + 3$ vectors in $\mathbb{R}^d$ for any $d \geq 5$ such that all vectors in the set have pairwise nonnegative dot product and that no orthogonal transformation can map all vectors of the set to the nonnegative orthant. Moreover, they showed that for any three vectors in $\mathbb{R}^3$ and any four vectors in $\mathbb{R}^4$ that have pairwise nonnegative dot product, there exists an orthogonal transformation that maps all vectors to the nonnegative orthant. Note that this gives a tight upper and lower bound for $\mathbb{R}^3$. We now give a tight upper bound for $\mathbb{R}^4$.

Proposition 1. *There is a set of five vectors in $\mathbb{R}^4$ with pairwise nonnegative dot product such that no orthogonal transformation can map all vectors in the nonnegative orthant.*

Proof. The idea of the proof is similar to the construction of Gray and Wilson [15] for $d \geq 5$. Consider the following five vectors:

$$\begin{aligned} v_1 &= (\ \ 1, \ \ 1, 0, 0\ ) \\ v_2 &= (\ \ 0, \ \ 0, 1, 0\ ) \\ v_3 &= (\ \ 0, \ \ 0, 0, 1\ ) \\ w_1 &= (-1, \ \ 2, 1, 3\ ) \\ w_2 &= (\ \ 2, -1, 1, 1\ ) \end{aligned}$$

Note that the five vectors indeed have pairwise nonnegative dot product. Moreover, v_1, v_2, v_3 are pairwise orthogonal and w_1 and w_2 are orthogonal. It is crucial to observe, however, that both w_1 and w_2 have positive dot product with each of v_1, v_2, v_3.

Suppose there is an orthogonal transformation T that maps v_1, v_2, v_3, w_1, w_2 to the nonnegative orthant. Recall that orthogonal transformations preserve the dot product between any two vectors. Hence, in particular, $T(v_1), T(v_2), T(v_3)$ are pairwise orthogonal, and $T(w_1)$ and $T(w_2)$ are orthogonal. Also note that all coordinates of $T(v_1), T(v_2), T(v_3), T(w_1), T(w_2)$ are nonnegative. If $T(v_i)$ and $T(v_j)$ are strictly positive in the same coordinate for $i, j \in \{1, 2, 3\}, i \neq j$, then their dot product is strictly positive, contradicting their orthogonality. Therefore, at least two of $T(v_1), T(v_2), T(v_3)$ span a coordinate axis. Without loss of generality, these are the third and fourth coordinate axes. Since T preserves the value of the dot product between any two vectors, it follows from the crucial observation above that $T(w_1)$ and $T(w_2)$ are both strictly positive in the third and fourth coordinate. As all coordinates of both $T(w_1)$ and $T(w_2)$ are nonnegative, the dot product of $T(w_1)$ and $T(w_2)$ is strictly positive, contradicting their orthogonality. Hence, T cannot exist. □