Discrete Mathematics 338 (2015) 291–305

Contents lists available at ScienceDirect

Discrete Mathematics

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

On the computation of edit distance functions

Ryan R. Martin

Department of Mathematics, Iowa State University, Ames, IA 50011, United States

ARTICLE INFO

Article history:
Received 16 December 2010
Received in revised form 16 September 2014
Accepted 16 September 2014
Available online 30 October 2014

Keywords:
Edit distance
Hereditary property
Symmetrization
Split graph
Colored regularity graph

ABSTRACT

The edit distance between two graphs on the same labeled vertex set is the size of the symmetric difference of the edge sets. The edit distance function of the hereditary property, $\mathcal{H}$, is a function of $p \in [0, 1]$ and is the limit of the maximum normalized distance between a graph of density p and $\mathcal{H}$.

This paper uses the symmetrization method of Sidorenko in order to compute the edit distance function of various hereditary properties. For any graph H, $\mathrm{Forb}(H)$ denotes the property of not having an induced copy of H. We compute the edit distance function for $\mathrm{Forb}(H)$, where H is any split graph, and the graph H_9, a graph first used to describe the difficulties in computing the edit distance function.

1. Introduction

For two graphs G and G' on the same labeled vertex set of size n, the *normalized edit distance* between them is denoted $\mathrm{dist}(G, G')$ and satisfies

$$\mathrm{dist}(G, G') = \left|E(G)\triangle E(G')\right| / \binom{n}{2}.$$

A *property* of graphs is simply a set of graphs. A *hereditary property* is a set of graphs that is closed under isomorphism and the taking of induced subgraphs. The normalized edit distance between a graph G and a property $\mathcal{H}$ is denoted $\mathrm{dist}(G, \mathcal{H})$ and satisfies

$$\mathrm{dist}(G, \mathcal{H}) = \min\left\{\mathrm{dist}(G, G') : V(G) = V(G'), G' \in \mathcal{H}\right\}.$$

In this paper, all properties will be hereditary.

1.1. The edit distance function

The *edit distance function* of a property $\mathcal{H}$, denoted $ed_{\mathcal{H}}(p)$, measures the maximum distance of a density-p graph from a hereditary property. Formally,

$$ed_{\mathcal{H}}(p) = \sup_{n\to\infty} \max\left\{\mathrm{dist}(G, \mathcal{H}) : |V(G)| = n, |E(G)| = \left\lfloor p\binom{n}{2}\right\rfloor\right\}.$$

Balogh and the author [8] use a result of Alon and Stav [3] to show that the supremum can be made into a limit, as long as the property $\mathcal{H}$ is hereditary.

$$ed_{\mathcal{H}}(p) = \lim_{n\to\infty} \max\left\{\mathrm{dist}(G, \mathcal{H}) : |V(G)| = n, |E(G)| = \left\lfloor p\binom{n}{2}\right\rfloor\right\}. \quad (1)$$

E-mail address: rymartin@iastate.edu.

http://dx.doi.org/10.1016/j.disc.2014.09.005
0012-365X/

Moreover, the result from [8] establishes that if $\mathcal{H}$ is hereditary then we also have

$$ed_{\mathcal{H}}(p) = \lim_{n\to\infty} \mathbb{E}\left[\mathrm{dist}(G(n,p),\mathcal{H})\right].$$

That is, the maximum edit distance to a hereditary property for a density-p graph is the same, asymptotically, as that of the Erdős–Rényi random graph $G(n,p)$ (see Chapter 10 of [1]).

For any nontrivial hereditary property $\mathcal{H}$ (that is, one that is not finite), the function $ed_{\mathcal{H}}(p)$ is continuous and concave down [8]. Hence, it achieves its maximum. The maximum value of $ed_{\mathcal{H}}(p)$ is denoted $d^*_{\mathcal{H}}$. The value of p at which this maximum occurs is denoted $p^*_{\mathcal{H}}$.

It should be noted that, for some hereditary properties, the edit distance function may achieve its maximum over a closed interval rather than a single point. In such cases, we will also let $p^*_{\mathcal{H}}$ denote the interval over which the given edit distance function achieves its maximum.

1.2. Symmetrization

In order to compute edit distance functions, we use the method of symmetrization, introduced by Sidorenko [15] and discussed in [12] as a way to compute edit distance functions. We will discuss what symmetrization is and how it is used in Section 4. It uses some properties of quadratic programming, first applied by Marchant and Thomason [11].

Some results on the edit distance function can be found in a variety of papers [14,6,7,3,2,5,4,11,13]. Much of the background to this paper can be found in a paper by Balogh and the author [8]. Terminology and proofs of supporting lemmas that are suppressed here can be found in [12].

1.3. Main results

Given a graph H, $\mathrm{Forb}(H)$ is the set of all graphs that have no induced copy of H. Clearly $\mathrm{Forb}(H)$ is a hereditary property for any graph H and such a property is called a *principal hereditary property*. It is easy to see that, for any hereditary property $\mathcal{H}$, there exists a family of graphs $\mathcal{F}(\mathcal{H})$ such that $\mathcal{H} = \bigcap_{H\in\mathcal{F}(\mathcal{H})} \mathrm{Forb}(H)$.

1.3.1. Split graphs

The main results of this paper are Theorems 1 and 3.

A *split graph* is a graph whose vertex set can be partitioned into one clique and one independent set. If H is a split graph on h vertices with independence number α and clique number ω, then $\alpha+\omega\in\{h,h+1\}$. The value of $p^*_{\mathrm{Forb}(H)}$ and of $d^*_{\mathrm{Forb}(H)}$ had been obtained for $H=K_{1,3}$, the claw, by Alon and Stav [2] and for graphs of the form K_a+E_b (an a-clique with b isolated vertices) by Balogh and the author [8].

For the $\mathrm{Forb}(K_a+E_b)$ result, the proof required a weighted version of Turán's theorem. The symmetrization method, however, is much more powerful and we can use it to obtain Theorem 1, which gives the value of the edit distance function for all $\mathrm{Forb}(H)$, where H is a split graph.

Theorem 1. *Let H be a split graph that is neither complete nor empty, with independence number α and clique number ω. Then,*

$$ed_{\mathrm{Forb}(H)}(p) = \min\left\{\frac{p}{\omega-1}, \frac{1-p}{\alpha-1}\right\}. \tag{2}$$

It is a trivial result (see, e.g., [12]) that $ed_{\mathrm{Forb}(K_\omega)}(p) = p/(\omega-1)$ and $ed_{\mathrm{Forb}(E_\alpha)}(p) = (1-p)/(\alpha-1)$. So, we know the edit distance function for all split graphs.

Corollary 2 follows immediately from Theorem 1 (and the following comment on trivial split graphs), giving the value of the maximum of the edit distance function and the value at which it occurs.

Corollary 2. *Let H be a split graph with independence number α and clique number ω. Then, $\left(p^*_{\mathcal{H}}, d^*_{\mathcal{H}}\right) = \left(\frac{\omega-1}{\alpha+\omega-2}, \frac{1}{\alpha+\omega-2}\right)$.*

To understand the importance of the upcoming Theorem 3, we must define the notion of colored regularity graphs.

1.3.2. Colored regularity graphs

If S and T are sets, then $S\dot\cup T$ denotes the disjoint union of S and T. If v and w are adjacent vertices in a graph, we denote the edge between them to be vw.

A *colored regularity graph (CRG)*, K, is a simple complete graph, together with a partition of the vertices into white and black $V(K) = \mathrm{VW}(K)\dot\cup\mathrm{VB}(K)$ and a partition of the edges into white, gray and black, $E(K) = \mathrm{EW}(K)\dot\cup\mathrm{EG}(K)\dot\cup\mathrm{EB}(K)$. We say that a graph H embeds in K, (writing $H\mapsto K$) if there is a function $\varphi : V(H)\to V(K)$ so that if $h_1h_2\in E(H)$, then either $\varphi(h_1)=\varphi(h_2)\in\mathrm{VB}(K)$ or $\varphi(h_1)\varphi(h_2)\in\mathrm{EB}(K)\cup\mathrm{EG}(K)$ and if $h_1h_2\notin E(H)$, then either $\varphi(h_1)=\varphi(h_2)\in\mathrm{VW}(K)$ or $\varphi(h_1)\varphi(h_2)\in\mathrm{EW}(K)\cup\mathrm{EG}(K)$.