Statistical Methodology 33 (2016) 31–54

Contents lists available at ScienceDirect

Statistical Methodology

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

Edge density of new graph types based on a random digraph family

Elvan Ceyhan

Department of Mathematics, Koç University, 34450 Sarıyer, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 31 March 2015
Received in revised form
3 April 2016
Accepted 14 July 2016
Available online 27 July 2016

MSC:
60D05
60F05
62E20
05C80
05C20
60C05

Keywords:
Arc density
Asymptotic normality
Central limit theorem
Delaunay tessellation and triangulation
Proximity catch digraph
Reflexivity
Underlying graph
U-statistic

ABSTRACT

We consider two types of graphs based on a family of proximity catch digraphs (PCDs) and study their edge density. In particular, the PCDs we use are a parameterized digraph family called proportional-edge (PE) PCDs and the two associated graph types are the "underlying graphs" and the newly introduced "reflexivity graphs" based on the PE-PCDs. These graphs are extensions of random geometric graphs where distance is replaced with a dissimilarity measure and the threshold is not fixed but depends on the location of the points. PCDs and the associated graphs are constructed based on data points from two classes, say $\mathcal{X}$ and $\mathcal{Y}$, where one class (say class $\mathcal{X}$) forms the vertices of the PCD and the Delaunay tessellation of the other class (i.e., class $\mathcal{Y}$) yields the (Delaunay) cells which serve as the support of class $\mathcal{X}$ points. We demonstrate that edge density of these graphs is a U-statistic, hence obtain the asymptotic normality of it for data from any distribution that satisfies mild regulatory conditions. The rate of convergence to asymptotic normality is sharper for the edge density of the reflexivity and underlying graphs compared to the arc density of the PE-PCDs. For uniform data in Euclidean plane where Delaunay cells are triangles, we demonstrate that the distribution of the edge density is geometry invariant (i.e., independent of the shape of the triangular support). We compute the explicit forms of the asymptotic normal distribution for uniform data in one Delaunay triangle in the Euclidean plane utilizing this geometry invariance property. We also provide various versions of edge density in the multiple triangle case. The approach presented here can also be extended for application to data in higher dimensions.

E-mail address: elceyhan@ku.edu.tr.

http://dx.doi.org/10.1016/j.stamet.2016.07.003
1572-3127/

1. Introduction

Proximity catch digraphs (PCDs) are a recently introduced digraph family and have applications in pattern classification and spatial data analysis. PCDs are random digraphs (i.e., directed graphs) in which each vertex corresponds to a data point, and arcs (i.e., directed edges) are defined in terms of some bivariate relation on the data. One type of PCD is the class cover catch digraph (CCCD) introduced by Priebe et al. [23] who gave the exact and the asymptotic distribution of its domination number for uniform data on bounded intervals in $\mathbb{R}$. Priebe et al. [25] and DeVinney and Priebe [13] applied the concept in higher dimensions and demonstrated relatively good performance of it in classification. Their methods involve *data reduction* (i.e., *condensing*) by using approximate minimum dominating sets as *prototype sets*, since finding the exact minimum dominating set is an NP-hard problem in general – e.g., for CCCDs in multiple dimensions – (see DeVinney and Priebe [13]). Our PCDs are constructed in a two-class setting with points from the class of interest (i.e., target class) that constitutes the vertices of the digraph. Let $\mathcal{X}_n$ and $\mathcal{Y}_m$ be two data sets of size n and m from classes $\mathcal{X}$ and $\mathcal{Y}$, respectively, and let class $\mathcal{X}$ be the target class. Then the vertices of the PCD are $\mathcal{X}_n$ and there is an arc from $x_1 \in \mathcal{X}_n$ to $x_2 \in \mathcal{X}_n$, based on a binary relation which measures the relative proximity of x_1 and x_2 with respect to the $\mathcal{Y}$ points. This relative proximity is determined based on the Delaunay tessellation of $\mathcal{Y}$ points. The PCDs are also closely related to the class cover problem of Cannon and Cowen [5] where the goal is finding a cover for the target class (i.e., finding a set of regions that contain all the points from the target class). Ceyhan and Priebe [8] introduced a digraph family called *proportional-edge PCD* (PE-PCD) and calculated the asymptotic distribution of its arc density and used it in spatial pattern testing [10]. PE-PCDs are parameterized digraphs with an expansion and a centrality parameter.

We consider two graph types defined based on the digraphs (in particular on PE-PCDs). The *underlying graph* of a digraph is obtained when any arc (between two vertices) is replaced by an edge disallowing multi-edges [11]. We introduce another graph type by replacing each symmetric or reflexive arc (between two vertices) with an edge and removing the (nonsymmetric) single arcs, and call it *reflexivity graph*. In a digraph $D = (V, A)$ with vertex set V and arc set A, an arc (a, b) is *symmetric* iff $\{(a, b), (b, a)\} \subset A$, i.e., the points a and b satisfy *reflexivity* with respect to the binary relation defining the arcs. The reflexivity and underlying graphs based on the PE-PCDs are also generalized versions of random geometric graphs (RGGs) (see, e.g., Penrose [22] for an extensive treatment of RGGs). Applications of RGGs include modeling/understanding disease spread among trees scattered in a forest, communication between a set of nests of animals or birds in a region of interest or between stations in a country or nerve cells in a living organism. RGGs are based on vertices independently and identically (iid) generated in $\mathbb{R}^k$ and an edge is inserted between two vertices if the distance between them does not exceed a certain threshold value. On the other hand, in our graphs, the regions that determine the edges would depend on the vertices, hence the threshold is adjusted based on the location of the vertices; and instead of a distance, a dissimilarity measure is employed. PCDs might also be applied in similar settings as those of RGGs, e.g., in testing spatial interaction (as in [10,9]); and in similar settings as those of the CCCDs in pattern classification (as in [24]).

We investigate the properties of the graph invariant, called *edge density*, of the reflexivity and underlying graphs of PCDs. Edge density of a graph is the ratio of number of edges to the total number of edges possible with the same set of vertices. Hence for a graph $G_n = (V, E)$ with vertex set of size $|V| = n$ and edge set E, edge density is $2|E|/(n(n-1))$. The maximal density is 1, which is attained for complete graphs, while the minimal density is 0, which is attained when $E = \emptyset$. The average degree of the graph G_n is defined as $2|E|/n$, is closely related to edge density; it is simply a scaled version of edge density. Edge density is also defined as $|E|/n$ by Grünbaum [14] who studies it for 4-critical planar graphs. In this article, we only use the quantity $2|E|/(n(n-1))$ as the edge density for the graph $G_n = (V, E)$. Arc density of a digraph $\mathbf{D}_n = (V, A)$ with $|V| = n$ is defined similarly as $|A|/(n(n-1))$. Edge density is also instrumental in determining the density of higher order structures in graphs, e.g., minimal density of triangles in graphs is provided in terms of edge density by Razborov [26]. Michael [21] studies the edge density for another special type of geometric graph called sphere of influence graph. A local version of edge density is also defined and investigated for subgraphs [20]. Furthermore, Darst et al. [12] define another local version called internal edge