U-statistics on network-structured data with kernels of degree larger than one

Yuyi Wang YUYI.WANG@CS.KULEUVEN.BE
Christos Pelekis CHRISTOS.PELEKIS@CS.KULEUVEN.BE
Jan Ramon JAN.RAMON@CS.KULEUVEN.BE
Computer Science Department, KU Leuven, Belgium

Editors: Wilhelmiina Hämäläinen, Francois Petitjean, and Geoff Webb

Abstract

Most analysis of U-statistics assumes that data points are independent or stationary. However, when we analyze network data, these two assumptions do not hold any more. We first define the problem of weighted U-statistics on networked data by extending previous work. We analyze their variance using Hoeffding's decomposition and also give exponential concentration inequalities. Two efficiently solvable linear programs are proposed to find estimators with minimum worst-case variance or with tighter concentration inequalities.

1. Introduction

Nowadays there is a plethora of real-world datasets which are network-structured. These are examples of relational databases, i.e. data samples are relations between objects, and so exhibit dependencies. A typical example is the web which, due to the explosion in social networks and the expansion of e-commerce, is generating an immense amount of network-structured data. Therefore we need statistical methods that permit us to mine and learn from this type of datasets. An example of a statistical method that generates unbiased estimators of minimum variance involves the notion of U-statistics. U-statistics are a class of measures, proposed by W. Hoeffding (Hoeffding, 1948), which can usually be written as averages over functions on elements or tuples of elements of samples, e.g., the sample mean, sample variance, sample moments, Kendall's τ (see (Kendall, 1938)), Wilcoxon's signed-rank sum (see (Wilcoxon, 1945)), etc.

Most analysis of U-statistics assumes that data points are independently distributed. However, when we consider networked data points, this assumption does not hold any more; two or more examples may share some common object.

In our previous work (Wang et al., 2014), we provided a statistical theory of learning from networked training examples. This work generalizes the results and extends the ideas further. A crucial assumption in our previous work was that every (perhaps correlated) data point is used only once. In contrast to this, U-statistics (see e.g. (Lee, 1990)) are a class of measures that allows us to repeatedly use data points. For example, the rank correlation estimator of Kendall (Kendall's τ) compares every data point to all other points. When we consider U-statistics on networked data points, data points are repeatedly used if the degree, d, of the kernel of U-statistics is greater than 1 (the case $d = 1$ has been discussed in (Wang et al., 2014)). Different data points may be also correlated. In this work we address the problem of how to design U-statistics, on networked data points, that exhibit small variance and small probability of deviation from their mean.

There is a vast literature on U-statistics for dependent random variables. However, most of the work focuses on providing central limit theorems and related results for dependent stationary sequences of random variables. For example, in (Khashimov, 1988; Hsing and Wu, 2004; Lee, 1990; Khashimov, 1994; Kim et al., 2011) the authors discuss U-statistics on several types of stationary sequences, like weakly dependent stationary sequences, m-dependent stationary sequences, absolutely regular process and random variables with mixing conditions, etc. The assumptions made in those works are not suitable for networked random variables which will be discussed in this paper. Our contribution is to not only analyze the variance and provide concentration bounds of U-statistics on networked random variables but also to design good U-statistics for this type of networked data.

In addition, there exists literature on weighted U-statistics. In (Ha et al., 2014), the authors analyze the asymptotic behavior of weighted U-statistics with i.i.d. data points. In (Nasari, 2012). the author considers incomplete U-statistics which are similar to our setting, but the attention is focused towards asymptotic results under the assumption of i.i.d. data points. In (O'Neil and Redner, 1993), it is shown that non-normal limits can occur for some choices of weights. In (Rifi and Utzet, 2000), one can find a sufficient condition for the convergence of weighted U-statistics. In (Hsing and Wu, 2004), the authors consider weighted U-statistics for stationary processes. Our results differ from the above in the fact that we do not assume independence and our attention is focused towards different aspects.

The rest of the paper is organized as follows. In Section 2, we define a weighted version of U-statistics on networked random variables and state the basic questions we are interested in. In Section 3, we bound the variance of the U-statistics by employing Hoeffding's decomposition. Subsequently, in Section 4, we formulate a linear program that allows us to obtain a concentration inequality for weighted U-statistics. In Section 5, we minimize the worst-case variance using a convex program. Finally, in Section 6, we draw conclusions with some remarks and comments on possible future work.

2. Preliminaries

In this section, we give a formal definition of the problem that is addressed in this paper. Let $G = (V^{(1)} \cup V^{(2)}, E)$ be a *bipartite* graph[1] and assume that we are given two sets of i.i.d. random variables that are indexed using the vertices of G. That is, let $\mathcal{X}^{(1)} = \{\phi_v\}_{v \in V^{(1)}}$ be a set of i.i.d., vector-valued random variables associated to $V^{(1)}$ and let $\mathcal{X}^{(2)} = \{\psi_v\}_{v \in V^{(2)}}$ be a set of i.i.d. random variables associated to $V^{(2)}$. Fix any enumeration $\{e_1, \ldots, e_n\}$ of the edge set E. To every edge $e_i = (u, v) \in E$, we associate a pair of random variables by setting $X_i = (\phi_v, \psi_u) \in \mathcal{X}^{(1)} \times \mathcal{X}^{(2)}$. We will denote by $X_i^{(1)}$ the first coordinate of X_i and by $X_i^{(2)}$ the second coordinate of X_i. Similarly, $e_j^{(i)}, i = 1, 2$, will denote the vertex of e_j that lies in $V^{(i)}$. We will refer to the set $X = \{X_i\}_{i=1}^n$ as a set of G-*networked random variables*. In addition, for $S \subseteq \{1, 2\}$, we will denote by $X_i^{(S)} = \times_{s \in S} X_i^{(s)}$ the (sub)vector formed by the coordinates of X_i that correspond to S. In particular $X_i^{(\emptyset)} = \emptyset$. For $S, T \subseteq \{1, 2\}$,

[1] We remark that our results can be extended to k-partite hypergraphs but, in order to keep the formalism simple, we present here the case $k = 2$.

we will denote by $X_i^{(S)} \cdot X_j^{(T)}$ the (sub)vector $Y \in \mathcal{X}^{(S \cup T)}$ for which $Y^{(S)} = X_i^{(S)}$ and $Y^{(T)} = X_j^{(T)}$.

Let $f(\cdot,\cdot)$ be a real valued function such that if e_i and e_j are disjoint edges in E (henceforth denoted $e_i \cap e_j = \emptyset$) then $\mathbb{E}[f(X_i, X_j)] = \mu$ and $\mathbb{E}\left[f^2(X_i, X_j)\right] - \mu^2 = \sigma^2$. Such a function $f(\cdot,\cdot)$ appears, for example, in the Kendall's τ rank correlation coefficient (see Example 1 below). Let us illustrate the above definitions with an example.

Example 1 *Let the vertex set $V^{(1)}$ represent a set of persons and $V^{(2)}$ represent a set of films. For every person $v \in V^{(1)}$ and every film $u \in V^{(2)}$ join the corresponding vertices with an edge if and only if person v has seen the film u. The result is a bipartite graph, $G = (V^{(1)} \cup V^{(2)}, E)$. An instance of such a graph can be found in Figure* 1. *Suppose that for every person $v \in V^{(1)}$ there is feature vector, ϕ_v, that contains information on, say, the gender, age, nationality, etc., of person v and that for every film $u \in V^{(2)}$ there is a feature vector, ψ_u, containing information on, say, scenography, actor popularity, etc., of film u. Thus, every edge $e_i = (v, u) \in E$ is associated to vector $X_i = (\phi_v, \psi_u)$. Now suppose that we have two functions, $S_1(\cdot), S_2(\cdot)$, that take values in $[0, 1]$ and are such that $S_k(X_i), k = 1, 2$, represents a rating/certificate that is given to a specific characteristic of film u by person v. If e_i, e_j are such that $e_i \cap e_j = \emptyset$, define the function $f(X_i, X_j)$ by setting*

$$f(X_i, X_j) = (-1)^{I\{S_1(X_i) > S_1(X_j)\} + I\{S_2(X_i) > S_2(X_j)\}},$$

where $I\{\cdot\}$ denotes indicator. Thus $f(X_i, X_j)$ is equal to 1 *if the ordering on both ratings agree and equal to* -1, *otherwise. Kendall's τ-coefficient (see (Kendall, 1938)) is defined as $\tau = \frac{2}{n(n-1)} \sum_{e_i, e_j} f(X_i, X_j)$, where the sum runs over all pairs of disjoint edges, e_i, e_j. Note that the fact that the function $f(\cdot,\cdot)$ is defined only for disjoint edges implies that τ is an unbiased estimator.* □

For a fixed bipartite graph, $G = (V, E)$, let us denote by $E^0 = \{(i, j) : e_i, e_j \in E \text{ and } e_i \cap e_j = \emptyset\}$ the set consisting of all pairs that are indices of disjoint edges from E (as an example, see Fig. 1. Suppose that we are given a function $w : E^0 \mapsto [0, +\infty)$ of nonnegative weights on the indices of pairs of disjoint edges from E. Set

$$U(f, w) = \frac{1}{|w|} \sum_{(i,j) \in E^0} w_{i,j} f(X_i, X_j), \tag{1}$$

where $|w| = \sum_{(i,j) \in E^0} w_{i,j}$. We will refer to $U(f, w)$ as a *weighted U-statistic* of f. Note that, by definition, $U(f, w)$ is an unbiased estimator of μ, or, more formally

$$\mathbb{E}[U(f, w)] = \mu = \mathbb{E}[f] \text{ for all } f, \tag{2}$$

and that, in order to guarantee this condition, it is important to sum over disjoint edges in Eq. (1). Hence the means of $U(f, w)$ and f are the same, but the same *may not* be true for the variance. Our attention in this paper (see Sections 3 and 5) is focused towards analyzing the variance of $U(f, w)$. Function $f(\cdot,\cdot)$ will be called the *kernel* of the U-statistic and will be considered as fixed throughout the paper; hence from now on we will denote $U(f, w)$ by $U(w)$. Because the kernel associates a real number to *two* vectors $X_i, X_j \in \mathcal{X}^{(1)} \times \mathcal{X}^{(2)}$, we say that its *degree* is two.

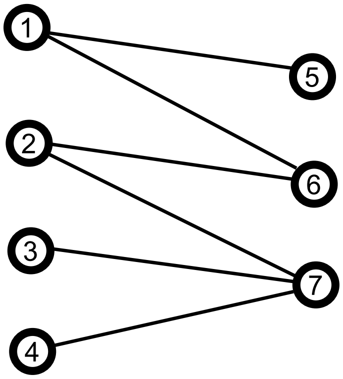

Figure 1: A bipartite graph. It contains nine pairs of disjoint edges: $(\{1,5\},\{2,6\})$, $(\{1,5\},\{2,7\})$, $(\{1,5\},\{3,7\})$, $(\{1,5\},\{4,7\})$, $(\{1,6\},\{2,7\})$, $(\{1,6\},\{3,7\})$, $(\{1,6\},\{4,7\})$, $(\{2,6\},\{3,7\})$, $(\{2,6\},\{4,7\})$. If the kernel is not symmetric then for each pair, say $(\{1,5\},\{2,6\})$, we have to include also the pair consisting of the same edges but written in reversed order, i.e., $(\{2,6\},\{1,5\})$.

In classical U-statistics (see (Hoeffding, 1948)), the variables $\{X_i\}_{i=1}^n$ are i.i.d. and all $w_{i,j}$ are equal to 1. By introducing weights in the above definition we will be able to obtain estimators that exhibit small variance and improved bounds on the probability of deviation from the mean.

Notice that the networked variables $\{X_i\}_{i=1}^n$ are *not* independent anymore, because two or more random variables may share the first or the second coordinate. For example, if $e_i^{(1)} = e_j^{(1)} = v$, then $X_i^{(1)} = X_j^{(1)} = \phi_v$.

In this paper we shall be interested in following basic questions:

- Can we find a sharp upper bound on the variance of $U(w)$?
- How can we bound the probability of deviation $\Pr[U(w) - \mu \geq t]$ for every fixed $t > 0$?
- How can we design a good (low variance and/or small probability of deviation) statistic $U(w)$ by suitably choosing the weight function w?

We investigate these questions in the subsequent sections.

3. Hoeffding's decomposition

In this section, we apply a technique, which is known as *Hoeffding's decomposition*, on weighted U-statistics of networked random variables. We begin by describing this well known technique (see (Hoeffding, 1948)).

Fix two independent random variables, say X_i and X_j, for which the corresponding edges are disjoint, i.e. $e_i \cap e_j = \emptyset$. For any two subsets $S, T \subseteq \{1,2\}$, we define $\mu_{(S,T)}\left(X_i^{(S)}, X_j^{(T)}\right)$ recursively via the following formula:

$$\mu_{(S,T)}\left(X_i^{(S)}, X_j^{(T)}\right) = \mathbb{E}\left[f(X_i, X_j) | X_i^{(S)}, X_j^{(T)}\right] - \sum_{(W,Z)\subset(S,T)} \mu_{(W,Z)}\left(X_i^{(W)}, X_j^{(Z)}\right)$$

where $(W,Z) \subset (S,T)$ means, by definition, $W \subseteq S$, $Z \subseteq T$ but $(W,Z) \neq (S,T)$ and $\mathbb{E}\left[f(X_i,X_j)|X_i^{(S)},X_j^{(T)}\right]$ denotes the conditional expectation of $f(X_i,X_j)$, given $X_i^{(S)}$ and $X_j^{(T)}$. Hoeffding's decomposition allows one to express $f(X_i,X_j)$ in terms of functions $\mu_{(S,T)}\left(X_i^{(S)},X_j^{(T)}\right)$, or, more formally

$$f(X_i,X_j) = \mathbb{E}[f(X_i,X_j)|X_i,X_j] = \sum_{S\subseteq\{1,2\},T\subseteq\{1,2\}} \mu_{(S,T)}\left(X_i^{(S)},X_j^{(T)}\right).$$

It is a well-known result, and in fact not so difficult to see, that the covariance of $\mu_{(S,T)}\left(X_i^{(S)},X_j^{(T)}\right)$ and $\mu_{(W,Z)}\left(X_i^{(W)},X_j^{(Z)}\right)$ is zero for $(S,T) \neq (W,Z)$, i.e., they are *uncorrelated*. This implies that

$$\sigma^2 = \sum_{S\subseteq\{1,2\},T\subseteq\{1,2\}} \sigma^2_{(S,T)} - \mu^2, \tag{3}$$

where $\sigma^2 = \mathbb{E}\left[f^2(X_i,X_j)\right] - \mu^2$ and $\sigma^2_{(S,T)} = \mathbb{E}\left[\mu^2_{(S,T)}\left(X_i^{(S)},X_j^{(T)}\right)\right]$. In other words, the variance of f can be partitioned into a sum of *variance-components*, where every component corresponds to a pair of subsets of $\{1,2\}$. Therefore, Hoeffding's decomposition allows us to write the function $f(X_i,X_j)$ as a sum of several *uncorrelated* functions.

This decomposition simplifies significantly the analysis of the variance of U-statistics based on an i.i.d. sample. To see this, let $\{X_i\}_{i=1}^n$ be i.i.d. and suppose that i,j,k are three different indices. Consider the U-statistics that are defined on $\{X_i\}_{i=1}^n$ with all weights being equal to 1. We want to find upper bounds on the variance of $U(w)$. Since the variance of $U(w)$ equals $\mathbb{E}\left[U(w)^2\right]-\mu^2$ (see also Eq. (1)), we have to find upper bounds on expressions of the form $\mathbb{E}[f(X_i,X_j)f(X_i,X_k)]$ and then add them up. Note that $\mathbb{E}[f(X_i,X_j)f(X_i,X_k)] - \mu^2$ is the *covariance* of $f(X_i,X_j)$ and $f(X_i,X_k)$. Now, in case one uses an i.i.d. sample, it can be shown that

$$\mathbb{E}[f(X_i,X_j)f(X_i,X_k)] - \mu^2 = \sigma^2_{(\{1\},\emptyset)} + \sigma^2_{(\{2\},\emptyset)} + \sigma^2_{(\{1,2\},\emptyset)}.$$

Thus, the variance of U decomposes into a sum of smaller variance-components. We remark that in the classical analysis of the variance of U-statistics using an i.i.d. sample we often assume that the kernel f is *symmetric*, i.e. $f(X_i,X_j) = f(X_j,X_i)$. The symmetry guarantees that the covariance of every possible pair, $f(X_i,X_j), f(X_m,X_l)$, can always be expressed as a sum of several variance-components.

However, the classical variance analysis using Hoeffding's decomposition cannot be directly applied to the case of networked random variables, due to dependence. To see this suppose that we have four different edges, say e_1,e_2,e_3,e_4, such that e_1 and e_3 intersect in $V^{(1)}$, i.e. $e_1^{(1)} = e_3^{(1)}$, and e_2 and e_3 intersect in $V^{(2)}$, i.e. $e_2^{(2)} = e_3^{(2)}$. Then, using the fact that the functions $\mu_{(\cdot,\cdot)}\left(X_i^{(\cdot)}, X_j^{(\cdot)}\right)$ are uncorrelated and some algebra, one can show that

$$\begin{aligned}
\mathbb{E}[f(X_1,X_2)f(X_3,X_4)] - \mu^2 &= \mathbb{E}\Big[\mu^2_{(\{1\},\emptyset)}(X_1^{(1)}) + \mu_{(\{2\},\emptyset)}(X_2^{(2)})\mu_{(\emptyset,\{2\})}(X_2^{(2)})\Big] \\
&+ \mathbb{E}\Big[\mu_{(\{1,2\},\emptyset)}(X_1^{(1)}\cdot X_2^{(2)})\mu_{(\{1\},\{2\})}(X_1^{(1)}, X_2^{(2)})\Big] \\
&= \sigma^2_{(\{1\},\emptyset)} + \mathbb{E}\Big[\mu_{(\{2\},\emptyset)}(X_2^{(2)})\mu_{(\emptyset,\{2\})}(X_2^{(2)})\Big] \\
&+ \mathbb{E}\Big[\mu_{(\{1,2\},\emptyset)}(X_1^{(1)}\cdot X_2^{(2)})\mu_{(\{1\},\{2\})}(X_1^{(1)}, X_2^{(2)})\Big],
\end{aligned}$$

Note that the second and the third term of the last expression do not decompose further to variance-components, i.e. into a sum of expressions of the form

$$\mathbb{E}\left[\mu^2_{(S,T)}\left(X_i^{(S)}, X_j^{(T)}\right)\right] = \sigma^2_{(S,T)}.$$

If we additionally assume that the kernel is symmetric, then the second term can be written in the form $\mathbb{E}\left[\mu_{(\{2\},\emptyset)}(X_2^{(2)})\mu_{(\emptyset,\{2\})}(X_2^{(2)})\right] = \sigma^2_{(\{2\},\emptyset)}$, but the third term can not.

Recall that we are interested in finding a sharp bound on the variance of U-statistics on networked examples. Recall further that the variance of weighted U-statistics is related to the covariance of $f(X_i,X_j)$ and $f(X_m,X_l)$, where $(e_i,e_j),(e_m,e_l) \in E^0$. In order to formally capture this relation, we will need the following definition.

Definition 1 (overlap index matrix) *Given a set of edges $E = \{e_i\}_{i=1}^n$ of a bipartite graph G, we define the overlap matrix of E, denoted J^E to be the $n\times n$ matrix whose (i,j) entry equals*

$$J^E_{i,j} = \{l \in \{1,2\} \mid e_i^{(l)} = e_j^{(l)}\}.$$

In other words, given two edges e_i, e_j from E, J^E_{ij} tells us the part of the graph on which they intersect. Note that $J^E_{i,j}$ is a subset of $\{1,2\}$. For example, in the graph of Fig. 1, if $e_1 = \{1,5\}$ and $e_2 = \{1,6\}$ then $J^E_{1,2} = \{1\}$, while if $e_1 = \{1,5\}$ and $e_3 = \{2,6\}$, then $J^E_{1,3} = \emptyset$.

If it is clear from the context, we will drop E from J^E and write J instead. Let $\{X_i\}_{i=1}^n$ be a set of G-networked random variables associated to $E = \{e_i\}_{i=1}^n$. Fix two pairs of edges, say (e_i,e_j) and (e_m,e_l), such that $e_i \cap e_j = \emptyset$ and $e_m \cap e_l = \emptyset$. One can show that the covariance of $f(X_i,X_j)$ and $f(X_m,X_l)$, i.e., the quantity $\Sigma(i,j;m,l) = \mathbb{E}[f(X_i,X_j)f(X_m,X_l)] - \mu^2$, is equal to

$$\sum_{*} \mathbb{E}\left[\mu_{(S\cup W,T\cup Z)}\left(X_i^{(S\cup W)}, X_j^{(T\cup Z)}\right)\mu_{(S\cup Z,T\cup W)}\left(X_i^{(S)}\cdot X_j^{(Z)}, X_i^{(T)}\cdot X_j^{(W)}\right)\right], \qquad (4)$$

where the sum $\sum_*$ runs over all quadruples (S,T,W,Z) such that $S \subseteq J_{i,m}, T \subseteq J_{j,l}, W \subseteq J_{i,l}, Z \subseteq J_{j,m}$.

Now, using the Cauchy–Schwarz inequality (the following expected values can be treated as inner products) it is easy to see that

$$\begin{aligned}
&\mathbb{E}\left[\mu_{(S\cup W,T\cup Z)}\left(X_1^{(S\cup W)}, X_2^{(T\cup Z)}\right)\mu_{(S\cup Z,T\cup W)}\left(X_1^{(S)}\cdot X_2^{(Z)}, X_2^{(T)}\cdot X_1^{(W)}\right)\right] \le \\
&\sqrt{\mathbb{E}\left[\mu^2_{(S\cup W,T\cup Z)}\left(X_1^{(S\cup W)}, X_2^{(T\cup Z)}\right)\right]}\sqrt{\mathbb{E}\left[\mu^2_{(S\cup Z,T\cup W)}\left(X_1^{(S)}\cdot X_2^{(Z)}, X_2^{(T)}\cdot X_1^{(W)}\right)\right]} = \\
&\sigma_{(S\cup W,T\cup Z)}\sigma_{(S\cup Z,T\cup W)}.
\end{aligned}$$

Summarizing, we can deduce the following bound on the variance of $U(w)$:

Theorem 2 *The variance of $U(w)$, i.e. the quantity $\mathbb{E}\left[U(w)^2\right] - \mu^2$, is at most*

$$\sum_{\odot} w_{i,j} w_{m,l} \sum_{*} \sigma_{(S\cup W, T\cup Z)} \sigma_{(S\cup Z, T\cup W)},$$

where $\sum_*$ is as before and $\sum_\odot$ runs over all quadruples (i,j,m,l) for which $e_i \cap e_j = \emptyset$ and $e_m \cap e_l = \emptyset$.

This bound is tight because it is possible to choose a kernel whose Hoeffding's decomposition ensures that the equality is attained in the Cauchy–Schwarz inequality, i.e. so that $\mu_{(S\cup W, T\cup Z)}$ and $\mu_{(S\cup Z, T\cup W)}$ are linearly dependent.

If we give every term $f(X_i, X_j)$ the same weight, the variance may not be minimal (see Section 5), and the same holds for the bound on the probability of deviation from the mean (see Section 4).

4. Concentration inequalities: a linear program method

In this section, we consider *bounded* kernels of degree two, i.e. functions, $f(\cdot,\cdot)$, that satisfy $|f - \mu| \leq M$, for some threshold $M > 0$. We are interested in obtaining concentration bounds for U-statistics with kernels of that form.

We would like to find a weight function for which the corresponding weighted U-statistics give sharp concentration inequalities

One way to get a bound is by applying Hoeffding's inequality as follows. Let us consider U-statistics that are based on a matching in the graph G. A *matching* in a hypergraph is a collection of disjoint edges and so, in the case of networked examples, it corresponds to an independent sample. If we use an independent sample of size α_G (the matching number of G), i.e., if we set

$$U_{ind} = \frac{1}{\binom{\alpha_G}{2}} \sum_{\{i,j \in E^* : i \neq j\}} f(X_i, X_j),$$

where E^* is a maximum matching of G, then by Hoeffding's result (see (Hoeffding, 1963, 1948)) we can conclude that if $\alpha_G \geq 2$, then

$$\Pr[U_{ind} - \mu \geq t] \leq \exp\left(-\frac{\alpha_G t^2}{4M^2}\right). \tag{5}$$

This bound may be sharp. However, it has two disadvantages:

1. It is difficult to find a large matching in a k-partite hypergraph when $k \geq 3$ (see (Garey and Johnson, 1979)), so the bound cannot be computed efficiently in more general graphs.

2. This method may lose some information of the sample since we remove some random variables from the sample.

Notice that finding a maximum matching in a hypergraph can be represented as an integer program. Integer programs are in general difficult to solve. In contrast to this, linear programs are much easier. With this in mind, and in order to avoid dealing with the aforementioned disadvantages, we formulate a linear program (LP) and use the solutions of the linear program as the weights of weighted U-statistics. This will require the notion of *vertex-bounded* weight function. For a given bipartite graph, $G = (V, E)$, recall the definition of the set $E^0 = \{(i,j) : e_i, e_j \in E \text{ and } e_i \cap e_j = \emptyset\}$

Definition 3 *A weight function w on E^0 is* vertex-bounded *if $w_{i,j} \geq 0$ for all pairs $(i,j) \in E^0$ and*

$$\textit{for all } v, \textit{ we have } \sum_{\{(i,j)\in E^0 : v\in e_i \textit{ or } v\in e_j\}} w_{i,j} \leq 1.$$

Our main result is the following concentration bound on vertex-bounded weighted U-statistics:

Theorem 4 *Let $X = \{X_i\}_{i=1}^n$ be a given set of networked random variables. If w is an vertex-bounded weight function, then the estimator $U(w)$ satisfies*

- *if $|f - \mu| \leq M$, then for any $t > 0$ we have*

$$\Pr[U(w) - \mu \geq t] \leq \exp\left(-\frac{|w|t^2}{2M^2}\right) \textit{ and } \tag{6}$$

- $\mathbb{E}\left[U(w)^2\right] - \mu^2 \leq \frac{\sigma^2}{|w|}$,

where $|w|$ is the sum of all weights $w_{i,j}$, with $(i,j) \in E^0$.

This theorem is an analogue of Theorems 18 and 23 in (Wang et al., 2014). Thus, in order to minimize the bounds of the previous theorem, one has to maximize $|w|$. This leads to the following linear program.

$$\begin{aligned} \max_w \quad & \sum_{i,j} w_{i,j} && (7) \\ \text{s.t.} \quad & \forall v : \sum_{\{(i,j)\in E^0 : v\in e_i \text{ or } v\in e_j\}} w_{i,j} \leq 1 && (8) \\ & \forall (i,j) \in E^0 : w_{i,j} \geq 0\ . && (9) \end{aligned}$$

We call the optimal objective value of this linear program the s'-value. Optimal weights w^* of this linear program will be referred to as s'-weights. Since the weight function corresponding to U_{ind} is vertex-bounded, it follows that $s' \geq \frac{\alpha_G}{2}$, when the matching number satisfies $\alpha_G \geq 2$. This shows that the bound given in Eq. (6) is smaller than the bound in Eq. (5). If the set of networked examples $\{X_i\}_{i=1}^n$ consists of i.i.d. random variables, then $s' = \frac{n}{2}$ provided $n \geq 2$. We remark that the bounds given in Theorem 4 have the advantage that the quantity s' can be computed efficiently, in polynomial time.

Note that the bounds in Theorem 4 depend on $|w|$ but they do not depend on function f. Note also that the first inequality of Theorem 4 is an analogue of the well known Hoeffding's inequality (see (Hoeffding, 1963)). In fact, using similar ideas, one can also show analogues of other well known exponential inequalities, e.g., Chernoff's or Bernstein's.

Now suppose that we use equal weights, i.e., we consider the following U-statistics:

$$U_{eqw} = \frac{1}{|E^0|} \sum_{(i,j)\in E^0} f(X_i, X_j).$$

Then we should replace the last constraint (9) with a constraint of the form:

$$w_{i,j} = t \geq 0, \text{ for all } (i,j) \in E^0. \tag{10}$$

Since we add more constraints to the LP, it follows that the optimal objective value of the new linear program will be smaller than the s'-value. This implies that the corresponding bounds on U_{eqw} cannot be smaller than those of an s'-weighted U-statistic. The following example shows that the difference between the optimal objective values may be large:

Example 2 *Consider the graph in Fig.* 1. *If we give the same weight to all pairs of disjoint edges, then $\sum_{i,j} w_{i,j} = \frac{9}{8}$. If we use an s'-weight function, then $\sum_{i,j} w_{i,j} = \frac{3}{2} > \frac{9}{8}$.*

The idea of using linear programs in order to obtain concentration bounds on sums of dependent random variables appears already in a paper by Svante Janson (Janson, 2004). However, Janson's bound involves the optimal objective value of a linear program that is known to be computationally hard. In (Wang et al., 2014) one can find concentration bounds on sums of network-structured random variables that improve Janson's bound and involve the optimal objective value of linear programs that can be computed efficently.

5. Minimum variance: a convex programming method

From the variance point of view, the s'-weight may not be the optimal solution. In this section we formulate a convex program which we use to minimize the worst-case variance of a U-statistics on a set of networked variables. To simplify our discussion, we only consider symmetric kernels and will provide remarks for the case of general kernels.

Given a bipartite graph, and using the version of Hoeffding's decomposition that has been described above, we see that the variance of $U(w)$ depends on the $2^4 - 2$ (because $\sigma_{(\emptyset,\emptyset)}$ does not affect and we fix the total variance σ) values of $\sigma_{(S,T)}$, one for each pair (S,T). Since we assume that the kernel is symmetric, two symmetric variance-components, e.g. $\sigma_{(\{1\},\emptyset)}$ and $\sigma_{(\emptyset,\{1\})}$, should be the same. In practice, we usually *do not* know the values of $\sigma_{(S,T)}$. Nevertheless, for every weight function w one can find a tight upper bound for $\text{var}(U(w))$ by maximizing $w^\top \Sigma w$ as a function of the variance components $\{\sigma_{(S,T)}\}_{S,T\subseteq\{1,2\}}$, where Σ is a covariance matrix defined by Eq. (4) (its row index is (i,j) and column index is (m,l)). We can see that when the structure, i.e. G, of networked random variables is given, the covariance matrix is determined by the values of $\sigma_{(S,T)}$. We will call a covariance matrix, Σ, for which $w^\top \Sigma w$ is maximum a *worst-case covariance matrix* and the corresponding variance $\text{var}(U(w))$ worst-case variance. A natural problem is to find the weight function, w, for which the worst-case variance is minimal. We do this by formulating a convex program. We begin with some lemmas that allow us to simplify this convex program.

Lemma 5 *For any fixed weight w, there exists some $\{\sigma^*_{(S,T)}\}_{S,T\subseteq\{1,2\}}$ which results in a worst-case covariance matrix (and equivalently worst-case variance) such that for all $S, T \subseteq \{1,2\}$ for which $|T|+|S| \geq 2$ we have $\sigma^*_{(S,T)} = 0$.*

The previous lemma holds true for non-symmetric kernels as well and should be compared with Lemma 16 in (Wang et al., 2014); its proof is also similar to the proof of Lemma 16. This result implies that we only need to consider worst-case covariance matrices for which all elements are zero except $\{\sigma_{(\{i\},\emptyset)}\}_{i\in\{1,2\}}$ and $\{\sigma_{(\emptyset,\{i\})}\}_{i\in\{1,2\}}$. Note that in case the kernel, f, is symmetric then we have $\sigma_{(\{i\},\emptyset)} = \sigma_{(\emptyset,\{i\})}$ for every $i \in \{1,2\}$. We can show one more lemma which simplifies further our problem.

Lemma 6 *Suppose that the weight function is fixed. If the kernel f is symmetric, then the worst-case variance is attained when $\sigma^2_{(\{q\},\emptyset)} = \sigma^2_{(\emptyset,\{q\})} = \frac{\sigma^2}{2}$ for some $q \in \{1,2\}$.*

For general kernel f, the worst-case variance-components can be attained by the method of Lagrange multipliers. Lemma 6 should also be compared with the remarks after Lemma 16 in (Wang et al., 2014).

Consequently, we can formulate the following optimization problem:

$$\begin{aligned}
\min_{w;t} \quad & t \\
\text{s.t.} \quad & \forall q \in \{1,2\}: \sum_{(i,j)\in E^0,(m,l)\in E^0} w_{i,j} w_{m,l} I_4 \leq t \\
& \sum_{i,j} w_{i,j} = 1 \\
& \forall i : w_{i,j} \geq 0,
\end{aligned}$$

where $I_4 = I\{q \in J_{i,m}\} + I\{q \in J_{i,l}\} + I\{q \in J_{j,m}\} + I\{q \in J_{j,l}\}$ and $I\{\cdot\}$ denotes the indicator function. This convex program is an analogue of program (7) in (Wang et al., 2014). Solving this convex quadratically constrained linear program, we can get weights which minimize the worst-case variance. Note that these weights may be not unique, but they form a convex region. By construction, these weights correspond to U-statistics whose variance is smaller than the variace of U-statistics corresponding to the s'-weight. If the variables $\{X_i\}_{i=1}^n$ are i.i.d. then optimal solutions of the above optimization problem satisfy $t = s' = \frac{n}{2}$, provided $n \geq 2$.

6. Conclusion

We considered the problem of how to analyze the quality of U-statistics on network data and how to design good estimators using weights. The analysis of variance based on Hoeffding's decomposition was generalized. We obtained a Hoeffding-type concentration bound for weighted U-statistics and, in order to minimize the bound, we used a linear program which can be solved efficiently. We also considered the worst-case variance, whose minimization results in a convex quadratically constrained linear program.

Though we only considered bipartite graphs and kernels of degree 2 in this paper, the results are valid for general k-partite hypergraphs and kernels of any degree d. A possible future work is to extend our results to V-statistics which is a class of biased estimators that are closely related to U-statistics.

Acknowledgements

This work was supported by the European Research Council, Starting Grant 240186 "MiGraNT: Mining Graphs and Networks, a Theory-based approach". We thank the anonymous referees for careful reading and suggestions that have improved the presentation.

References

Michael R. Garey and David S. Johnson. *Computers and intractibility, a guide to the theory of NP-Completeness.* W. H. Freeman Company, 1979.

Hyung-Tae Ha, Mei Ling Huang, and De Li Li. A remark on strong law of large numbers for weighted U-statistics. *Acta Math. Sin. (Engl. Ser.)*, 30(9):1595–1605, 2014. ISSN 1439-8516.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58.301:13–30, 1963.

Tailen Hsing and Wei Biao Wu. On weighted U-statistics for stationary processes. *The Annals of Probability*, 32(2), 2004.

Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24.3:234–248, 2004.

Maurice. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938.

Sh A Khashimov. On the Asymptotic Distribution of the Generalized U-Statistics for Dependent Variables. *Theory of Probability & Its Applications*, 32(2):373–375, 1988.

Sh A Khashimov. The central limit theorem for generalized U-statistics for weakly dependent vectors. *Theory of Probability & Its Applications*, 38(3):456–469, 1994.

Tae Yoon Kim, Zhi-Ming Luo, and Chiho Kim. The central limit theorem for degenerate variable U-statistics under dependence. *Journal of Nonparametric Statistics*, 23(3):683–699, 2011.

Justin Lee. *U-statistics: Theory and Practice.* CRC Press, 1990.

Masoud M Nasari. Strong law of large numbers for weighted U-statistics: Application to incomplete U-statistics. *Statistics & Probability Letters*, 82(6):1208–1217, 2012.

Kevin A. O'Neil and Richard A. Redner. Asymptotic Distributions of Weighted U-Statistics of Degree 2. *The Annals of Probability*, 21(2):1159–1169, 1993.

Mohamed Rifi and Frederic Utzet. On the Asymptotic Behavior of Weighted U-Statistics. *Journal of Theoretical Probability*, 13(1):141–167, 2000.

Yuyi Wang, Jan Ramon, and Zheng-Chu Guo. Learning from networked examples. *submitted to Journal of Machine Learning Research*, 2014.

Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6): 80–83, 1945.