Discovering Interesting Cycles in Graphs

Florian Adriaens, Jefrey Lijffijt, Tijl De Bie
IDLab, Ghent University
first.last@ugent.be

Cigdem Aslay, Aristides Gionis
Dept. of Computer Science, Aalto University
first.last@aalto.fi

ABSTRACT

Cycles in graphs often signify interesting processes. For example, cyclic trading patterns can indicate inefficiencies or economic dependencies in trade networks, cycles in food webs can identify fragile dependencies in ecosystems, and cycles in financial transaction networks can be an indication of money laundering. Identifying such interesting cycles, which can also be constrained to contain a given set of query nodes, although not extensively studied, is thus a problem of considerable importance.

In this paper, we introduce the problem of discovering interesting cycles in graphs. We first address the problem of quantifying the extent to which a given cycle is interesting for a particular analyst. We then show that finding cycles according to this interestingness measure is related to the Longest cycle and Maximum Mean weight cycle problems (in the unconstrained setting) and to the Maximum Steiner cycle and Maximum Mean Steiner cycle problems (in the constrained setting).

We show that the problems of finding the most interesting cycle and Steiner cycle are both **NP**-hard, and are **NP**-hard to approximate within a constant factor in the unconstrained setting, and within a factor polynomial in the input size for the constrained setting. We also show that the latter inapproximability result implies a similar result for the Maximum Steiner cycle and Maximum Mean Steiner cycle problems. Motivated by these hardness results, we propose a number of efficient heuristic algorithms and demonstrate their practical utility on a real-world use case.

KEYWORDS

Subjective interestingness, Minimum mean weight cycle, Maximum mean Steiner cycle

ACM Reference Format:
Florian Adriaens, Jefrey Lijffijt, Tijl De Bie and Cigdem Aslay, Aristides Gionis. 2019. Discovering Interesting Cycles in Graphs. In *MLG2019: 15TH INTERNATIONAL WORKSHOP ON MINING AND LEARNING WITH GRAPHS, August 5, 2019, Anchorage, Alaska - USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cycles occur as a natural data-mining pattern in several real-world applications. They appear naturally in food webs, where cycles highlight cyclic dependencies, often revealing the fragile parts of an ecosystem [8]. In financial transaction data, a cycle could be an indication of a money-laundering scheme [3]. In biological and complex networks, a cycle is an indication of a feedback mechanism [16]. Despite the wide range of use cases, the problem of discovering cyclic patterns in graphs has not received much attention in the data-mining community (See Section 7).

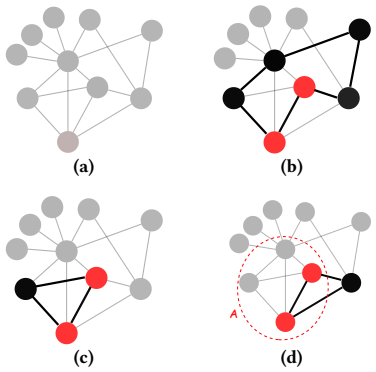

Figure 1: The most interesting Steiner cycles connecting the red nodes according to different prior beliefs on the graph shown in (a): when we have (b) no knowledge about the graph, (c) knowledge about the individual degrees, and (d) knowledge about the degrees and the density of the community $\mathcal{A}$.

In this paper, we study the problem of discovering interesting cycles in a directed and non-negatively weighted graph. We also consider the constrained case, where the cycles have to contain a set of user-specified query nodes. Cycles containing a given set of query nodes are called *Steiner* cycles [20, 22]. Identifying interesting Steiner cycles can be particularly useful in different application domains. For example, a biologist may be interested in finding a food chain that contains both a rabbit and a hawk to assess the importance of the hawk population for the rabbit population. Economists may be interested in finding surprising trading action between certain countries in different parts of the world.

As networks typically contain numerous cycles, a key challenge is the choice of a suitable interestingness measure for a cycle. It is clear that such a measure has to be subjective, i.e., taking into account which network characteristics (if any) are known a priori to the analyst. For example, for a lay person it might be surprising that more than 50% of the Dominican Republic's export is to the USA.[1] However, for an economist possessing the knowledge that those countries have a bilateral trade agreement (which can be formalized as prior information on the trade network), such a trade volume might not come as a surprise. Thus, we are interested in

MLG2019, August 5, 2019, Anchorage, Alaska - USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://tradingeconomics.com/dominican-republic/exports-by-country

designing methods that are able to take such prior knowledge into account.

Based on this observation, our proposed measure is built on *subjective interestingness*, a concept that was introduced by Silberschatz and Tuzhilin [21], and later extended by De Bie [6]. In the formalization of subjective interestingness, a pattern is deemed interesting if it is both *surprising* to the user and can be communicated in a *concise* way.

Figure 1 illustrates an example of our setting. Figure 1a shows a toy graph in which a user wishes to find a cyclic pattern containing the red query nodes. We consider three different users. The first user has no knowledge of the graph. For every revealed edge, the user learns something about the graph, hence, the most interesting cycle is the longest cycle containing the red nodes, as shown in Figure 1b. The second user has knowledge about the degrees of each node in the network. In this case, edges containing high degree nodes are less interesting to this user, as they are expected. This prior knowledge makes the cycle shown in Figure 1c the most interesting cycle. Our last user is a specialist. Besides knowing the degrees of the nodes, he also has prior knowledge that the red nodes are part of a dense community $\mathcal{A}$. Intra-community edges are now expected and thus are less interesting. This makes the cycle obtained in Figure 1d the most interesting cycle for the third user.

Following the proposed cycle interestingness measure, in Section 3 we formally define the two problem variants that we study in this paper: (i) the Maximum Subjectively Interesting Cycle problem (MSIC$[\alpha, \beta]$); and (ii) the Maximum Subjectively Interesting Steiner Cycle problem (k-MSIC$[\alpha, \beta]$), in which the cycle is required to contain a given set of k terminal vertices. We provide an extensive computational complexity analysis in Section 4 showing that both problems are **NP**-hard, and are **NP**-hard to approximate within a constant factor for MSIC$[\alpha, \beta]$, and within a factor polynomial in the input size for k-MSIC$[\alpha, \beta]$. We also show that the latter inapproximability result implies a similar result for the *Maximum Steiner cycle* and the *Maximum mean Steiner cycle* problems. The paper is organized as follows:

Contributions and roadmap.

- We present a novel subjective interestingness measure for cycle patterns in directed graphs (Section 2).
- We formally define the Maximum Subjectively Interesting Cycle and Maximum Subjectively Interesting Steiner Cycle problems, and provide an extensive theoretical analysis of their computational complexity (Sections 3 and 4).
- We propose a number of efficient and effective heuristics for both problems (Section 5).
- We experimentally verify the effectiveness of our methods and demonstrate their practical utility on a real-world use case: financial trade data between countries (Section 6).

2 CYCLES AND THEIR INTERESTINGNESS

In this section, we first introduce the notation used in this paper and formally define the notion of a cycle pattern in weighted digraphs (Section 2.2). We then explain how the interestingness of a cycle pattern can be formalized w.r.t. a background distribution that models prior knowledge about the structure of the graph (Section 2.3). For the sake of clarity and completeness, we also briefly summarize the related work on how such a background distribution can be derived based on a number of relevant types of prior knowledge on the graph structure (Section 2.4).

2.1 Graph notation

We assume a simple digraph $G = (V, E)$, with $|V| = n$ vertices and $|E| = m$ directed edges. A *walk* in G is defined as a sequence $v_1, v_2, \ldots, v_k$ of vertices, where $(v_i, v_{i+1}) \in E$ for $i \in [1, k-1]$ and $(v_i, v_{i+1}) \neq (v_j, v_{j+1})$ for all $1 \leq i < j \leq k$. We say that a walk is *closed* if $v_1 = v_k$. A (simple) *cycle* is a closed walk $v_1, v_2, \ldots, v_k = v_1$, with no repetition of the vertices v_i, for $1 < i < k$. We use $v \in C$ and $e \in C$ to indicate that a vertex v and an edge e is part of a cycle C, respectively. We use $|C|$ to denote the length of a cycle C, i.e. the number of edges it contains.

2.2 Cycles as patterns

The patterns considered in this paper consist of the specification of a cycle C that is stated to be present in a given graph. Additionally, we communicate $|C|$ positive real values ℓ_e, one for each edge in $e \in C$. Each value ℓ_e represents a lower bound[2] on the weight of edge e, thus informing the user that the weight is at least ℓ_e. In practice, in the most interesting cycle patterns, a lower bound ℓ_e will be equal (or as close as possible given the number encoding used) to the observed value of the weight $\mu(e)$, as a larger ℓ_e provides more information.

2.3 Subjective interestingness of cycle patterns

We follow the approach proposed by De Bie [6] in formalizing the subjective interestingness of a cycle pattern as the ratio of its *Information Content* (IC), and its *Description Length* (DL), which should reflect the amount of effort it takes the data analyst to assimilate the pattern. Here, IC is the negative log probability of the pattern being present in the data, where the probability is computed w.r.t. a so-called *background distribution* P which represents the prior expectations of the analyst. The IC reflects the fact that the more improbable the analyst considers a given pattern, the more information it conveys when the analyst learns the pattern is actually present.

It may be impossible to accurately represent all expectations of an analyst in a background distribution. Yet, it was argued that given a set of constraints in terms of expectations on certain statistics of the network (e.g., node degrees, subgraph densities, etc.), a robust estimate of the background distribution can be obtained by choosing P as the maximum entropy distribution, subject to these constraints [6].

As reviewed in Section 2.4, a wide range of prior knowledge types have the (convenient) property that the resulting background distribution factorizes as a product of independent distributions, one for each possible edge $e \in V \times V$. Hence, the IC of a cycle C

[2] We use a lowerbound and not the actual edge value. Often a user will not care about the *exact* actual weight of an edge, but rather only if the weight is high or not, relative to the user's expectations on the graph structure.

equals

$$\mathrm{IC}(C) = -\log\left(\prod_{e\in C} \Pr(\mu(e) \geq \ell_e)\right) = \sum_{e\in C} w(e), \tag{1}$$

where $w(e) \triangleq -\log(\Pr(\mu(e) \geq \ell_e))$ denotes the information content of the edge e, with $\Pr(\cdot)$ denoting the probability under the background distribution P. Note that $w(e) \geq 0$ for all $e \in V \times V$.

The DL can be computed similarly as in [17]. To communicate a cycle pattern C to the user, we need to communicate $|C|$ nodes. We assume that the cost of assimilating that a vertex is part of C is $\log(1/q)$, and that a vertex is not part of C is $\log(1/(1-q))$. Hence the DL of communicating $|C|$ nodes is equal to

$$|C| \cdot \log\frac{1}{q} + (n - |C|) \cdot \log\frac{1}{1-q} = |C| \cdot \log\frac{1-q}{q} + n \cdot \log\frac{1}{1-q},$$

for $0 < q < 1/2$. Here, q can be loosely interpreted as the expected probability that a random vertex is part of C, according to the user. Typically, q is to be chosen small. Hence, the DL of a cycle pattern equals

$$\mathrm{DL}(C) = \alpha|C| + n\beta.$$

where $\alpha > 0$ and $\beta > 0$ are defined as

$$\alpha = \log\frac{1-q}{q}, \beta = \log\frac{1}{1-q}. \tag{2}$$

We now formally define the subjective interestingess of a cycle pattern.

Definition 1 (Subjective Interestingness). *Given a directed graph $G = (V, E)$ with non-negative edge weights w, and parameters $\alpha > 0$ and $\beta > 0$, the subjective interestingness $F(C)$ of a cycle C is defined*[3] *as:*

$$F(C) = \frac{IC(C)}{DL(C)} = \frac{\sum_{e\in C} w(e)}{\alpha|C| + n\beta}. \tag{3}$$

2.4 Modeling a user's prior beliefs

As argued by De Bie [6], a good choice for the background distribution P is the maximum entropy distribution, subject to particular user expectations as linear constraints on the distribution. Here, the domain of the distribution P is the set of all possible edges over a given set of vertices. For a better understanding of these models, we recap some existing results and discuss a toyexample below.

2.4.1 A prior on the weighted in- and out-degrees. In the case of a prior belief on the weighted in- and out-degree of each node, the distribution P factorizes as a product of independent geometric distributions, one for each node pair. As discussed in [7], using a background distribution with the empirically weighted in- and out-degrees as constraints will ensure that cycle patterns are more interesting if they involve edges from low out-degree nodes to low in-degree nodes, ceteris paribus. As it is quite common that weighted node degrees are well-understood (e.g., biologists have a good idea about the predatory component of the diet of different species in a food web), this is an important type of background distribution in practice.

[3]We note that $F(\emptyset) = 0$.

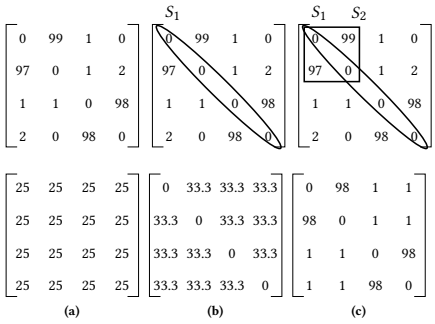

Figure 2: (top row) A toy graph, with constraints on the in- and out degrees of each node (a, b and c), combined with constraints on the densities of the sets S_1 (b and c) and S_2 (c). (bottom row) The expected values of the edges according to the MaxEnt distribution.

2.4.2 Additional priors on the density of any subsets. Additionally, extra constraints on the density on a number of user-provided subgraphs can be incorporated. For example, an economist might have knowledge of high trading volume between a group of neighboring countries, e.g., due to a free trade agreement or a common market, or a user might know that no self-edges exist in a network. In this case, if an edge $e = (i, j)$ is part of the specified subgraph, the probability that this edge has a weight at least ℓ becomes:

$$\Pr(\mu(e) \geq \ell) = \exp\big(-\ell(\lambda_i^{\text{out}} + \lambda_j^{\text{in}} + \lambda^{\text{block}})\big), \tag{4}$$

where $\lambda_i^{\text{out}} + \lambda_j^{\text{in}} + \lambda^{\text{block}} > 0$. Here, λ_i^{out} and λ_j^{in} denote the Lagrange multipliers associated with the resp. row- and column sums of node i and j, and λ^{block} denotes the Lagrange multiplier associated with the density of the specified subgraph. Adriaens et al. [1] showed how these multipliers can still be computed efficiently for large sparse networks and a limited number of specified subgraphs. Figure 2 shows an example of fitting the MaxEnt model on a 4x4 adjacency matrix $\mathbf{A}$ with different types of constraints. It illustrates how adding more constraints results in a closer fit of the background distribution to the empirical network. The probability $\Pr(\mathbf{A}_{12} \geq 99) = 0.038$ for (a), 0.054 for (b) and 0.53 for (c).

3 PROBLEM DEFINITION

The first problem considered in this paper is the problem of finding the "Maximum Subjectively Interesting Cycle" in a graph.[4] Formally:

Problem 1 (MSIC[α, β]). *Given a directed graph G with non-negative edge weights, and parameters $\alpha, \beta > 0$, find a simple cycle C such that $F(C)$ is maximized.*

Moreover, we can constrain the cycle to include a given set of terminal vertices to find "Maximum Subjectively Interesting Steiner Cycle". This leads to the second problem we address in this paper:

[4]Note that although the problem appears to have two parameters, in reality this can be reduced to one, e.g. by multiplying the objective with β and substituting α/β with a single parameter γ.

Problem 2 (k-MSIC$[\alpha, \beta]$). *Given a directed graph G with non-negative edge weights, a set of k terminal vertices, and parameters $\alpha, \beta > 0$, find a simple cycle C such that C contains all the terminals and $F(C)$ is maximized.*

MSIC$[\alpha, \beta]$ is closely related to two well-known graph problems. For $\alpha = 0$, MSIC$[\alpha, \beta]$ is equivalent to the problem of finding the *longest cycle in a digraph*, an **NP**-hard problem that is known for its difficulty to approximate [2, 9]. On the other hand, for $\beta = 0$, MSIC$[\alpha, \beta]$ is equivalent to the problem of finding a *maximum mean-cycle* in a directed graph with non-negative edge weights. This problem can be solved in polynomial time by using Karp's algorithm [14]. Although MSIC$[\alpha, \beta]$ is closely related to a tractable and to an **NP**-hard problem, it is not equivalent to either one as our problem setting assumes $\alpha > 0$ and $\beta > 0$. Yet, in Section 4 we show that MSIC$[\alpha, \beta]$ is **NP**-hard (as the longest cycle problem), while we discuss how Karp's algorithm can be used to provide approximations. This is a plausible approach, as in practice $\alpha \gg \beta$ (it takes more effort to assimilate the fact that a node is part of a cycle pattern than that a node is not part of a cycle pattern), such that the interestingness measure is closer to the maximum mean-cycle objective than to the longest cycle objective.

Likewise, k-MSIC$[\alpha, \beta]$ is closely related to two Steiner cycle problem variants. For $\alpha = 0$, k-MSIC$[\alpha, \beta]$ is equivalent to the problem of finding a *Maximum Steiner cycle*, i.e. Steiner cycle with maximum total weight, and for $\beta = 0$, k-MSIC$[\alpha, \beta]$ is equivalent to the problem of finding a *Maximum Mean Steiner cycle* in a digraph with non-negative edge weights. To the best of our knowledge, there are no known results on the approximability of both problems. Besides being **NP**-hard, we show in the next section that neither of these Steiner cycle problems, nor k-MSIC$[\alpha, \beta]$, can be approximated within a ratio that is polynomial in the number of vertices.

4 COMPUTATIONAL COMPLEXITY

The **NP**-hardness of both MSIC$[\alpha, \beta]$ and k-MSIC$[\alpha, \beta]$ follows directly from Lemma 1 and 3. Both lemmas, together with Lemma 2, are dedicated to showing inapproximability results for both problems. We note that the reduction in Lemma 3 can directly be applied to the Max. Steiner Cycle and Max. Mean Steiner Cycle problems, which is a novel result in itself.

Lemma 1. *There exists no constant-factor polynomial-time approximation algorithm for MSIC$[\alpha, \beta]$, unless* **P** = **NP**.

Proof. To prove this, we use an approximation preserving reduction from the Longest Cycle problem in digraphs [2]. Specifically, we use an *A-reduction* [4] that preserves membership in **APX**, which is the class of **NP** optimization problems that admit polynomial-time constant-factor approximation algorithms.

To show that a reduction from Longest Cycle problem to MSIC$[\alpha, \beta]$ is an *A-reduction*, we need to show that (i) there exists a polynomial-time computable function g mapping the solutions of MSIC$[\alpha, \beta]$ to the solutions of the Longest Cycle problem, and (ii) a polynomial-time computable function $c : Q \cap (1, \infty) \to Q \cap (1, \infty)$ such that any algorithm providing r-approximation to MSIC$[\alpha, \beta]$ with the approximate solution C provides $c(r)$-approximation to the Longest Cycle problem using the approximate solution $g(C)$.

Let $G = (V, E)$ be a given an instance of the Longest Cycle problem in digraphs. We construct an instance of MSIC$[\alpha, \beta]$ by assigning a constant weight $w(e) = \rho$, $\forall e \in E$ in G. Assume there exists a polynomial-time algorithm A which provides r-approximation to MSIC$[\alpha, \beta]$ for some constant $r \geq 1$. Let C^* denote the optimal solution to MSIC$[\alpha, \beta]$ and let C_A denote the solution returned by algorithm A. Then we have,

$$\frac{F(C^*)}{F(C_A)} = \frac{\rho|C^*|}{\rho|C_A|} \cdot \frac{\alpha|C_A| + n\beta}{\alpha|C^*| + n\beta} \leq r \tag{5}$$

Reminding that $F(C)$ monotonically increases with $|C|$ in such instances of MSIC$[\alpha, \beta]$ with uniform edge weights, we define g as the identity function, and use the solutions of MSIC$[\alpha, \beta]$ as the solutions of the Longest Cycle problem. Then, by re-arranging Eq.5 and using the fact that $2 \leq |C| \leq n$ for any cycle C, we have:

$$\begin{aligned} \frac{|C^*|}{|C_A|} &\leq r \cdot \frac{\alpha|C^*| + n\beta}{\alpha|C_A| + n\beta} \\ &\leq r \cdot \frac{n(\alpha + \beta)}{2\alpha + n\beta} \\ &\leq r \cdot (1 + \alpha/\beta) \end{aligned}$$

We have just showed that the Longest Cycle problem is A-reducible to MSIC$[\alpha, \beta]$. Finally, given that the Longest Cycle problem in digraphs is not in **APX** [2, 9], we conclude that MSIC$[\alpha, \beta]$ is also not in **APX**. □

Björklund et al. [2] show that there exists no polynomial-time approximation algorithm for the Longest Cycle problem in unweighted Hamiltonian digraphs with performance ratio $n^{1-\epsilon}$ for any fixed $\epsilon > 0$, unless **P** = **NP**. Next we show the implications of this strong inapproximability result for solving MSIC$[\alpha, \beta]$ in Hamiltonian digraphs with uniform edge weights.

Lemma 2. *It is* **NP***-hard to approximate MSIC$[\alpha, \beta]$ in a Hamiltonian digraph with uniform weights within a factor of*

$$\frac{n^{1-\epsilon} + \alpha/\beta}{1 + \alpha/\beta},$$

for any $\epsilon > 0$, unless **P** = **NP**.

Proof. Let $G = (V, E)$ be an unweighted Hamiltonian digraph denoting an instance of the Longest Cycle problem[2]. Given $G = (V, E)$, we construct an instance of MSIC$[\alpha, \beta]$ by assigning a constant weight to every edge, $w(e) = \rho$, $\forall e \in E$. Assume by contradiction that there exists such an approximation algorithm A which finds a solution C_A satisfying

$$\frac{\rho|C_A|}{\alpha|C_A| + n\beta} \geq \frac{1 + \alpha/\beta}{n^{1-\epsilon} + \alpha/\beta} \cdot \frac{\rho n}{\alpha n + n\beta} \tag{6}$$

By re-arranging the terms in Eq.6, we obtain $|C_A| \geq n^{\epsilon}$ implying that any such approximation algorithm to MSIC$[\alpha, \beta]$ leads to a polynomial-time $n^{1-\epsilon}$-approximation algorithm for the Longest Cycle problem in unweighted Hamiltonian digraphs, which is a contradiction, unless **P** = **NP**. □

Next we show the hardness of approximating k-MSIC$[\alpha, \beta]$.

Lemma 3. *It is* **NP***-hard to approximate k-MSIC$[\alpha, \beta]$ within a factor polynomial in the input size in digraphs with non-negative edge weights for any $k \geq 1$, unless* **P** = **NP**.

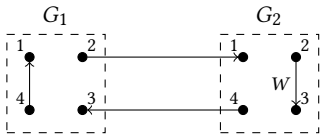

Figure 3: A visualization of the construction in Lemma 3.

Proof. To prove this, we use a reduction from the **NP**-complete Restricted Two Vertex Disjoint Paths problem (R2VDP), which was introduced by Björklund et al. [2] as the restricted version of the Two Vertex Disjoint Paths problem (2VDP) [19]. Given a digraph of order $n \geq 4$ and four vertices, 2VDP problem seeks to determine whether there exist two vertex disjoint paths, one from vertex 1 to 2 and one from vertex 3 to 4. In the restricted version R2VDP of 2VDP, all the YES instances of 2VDP are guaranteed to contain two such paths that together exhaust all vertices of G, i.e., the graph G with the additional edges from vertex 2 to 3 and from 4 to 1, contains a Hamiltonian cycle through these edges in YES instances to R2VDP.

Assume that there exists an approximation algorithm for k-MSIC$[\alpha, \beta]$ with ratio $p(n) \geq 1$ that is a polynomial of n. We show how to decide R2VDP by using such algorithm with approximation ratio $p(n)$. Given an instance of R2VDP, we construct an instance of k-MSIC$[\alpha, \beta]$ as follows. We connect 2 copies G_1 and G_2 of G by adding edges (i) from vertex 2 in G_1 to vertex 1 in G_2, and (ii) from vertex 4 in G_2 to vertex 3 in G_1. We also add an edge $(4, 1)$ in G_1 and an edge $(2, 3)$ in G_2. For each edge we assign a weight of 1, except for the edge $(2, 3)$ in G_2 for which we assign a weight of $W = n \cdot p(n) + 1$. Finally, we set the vertex 1 of G_1 as the terminal for 1-MSIC$[\alpha, \beta]$. Let $G' = (V', E')$ denote the resulting graph, as shown in Figure 3.

Let C^* denote the optimal solution to 1-MSIC$[\alpha, \beta]$ in G'. If G is a YES instance of R2VDP, then C^* is a Hamiltonian cycle in G', containing $2n$ edges with a total weight of $2n + n \cdot p(n)$, since, $F(C^*) = \dfrac{2n + n \cdot p(n)}{2\alpha n + n\beta} > \dfrac{|C|}{\alpha|C| + n\beta}$ for any other Steiner cycle C that is not Hamiltonian, thus, not containing the edge $(2, 3)$ in G_2. On the other hand, if G is a NO instance to R2VDP, then C^* can have at most $2n - 2$ edges, excluding the edge $(4, 1)$ in G_1 and the edge $(2, 3)$ in G_2, thus, $\dfrac{|C^*|}{|C^*| + 1} \leq \dfrac{2n - 2}{\alpha(2n - 2) + n\beta}$.

We have just shown that, unless **P** = **NP**, it is not possible to approximate 1-MSIC$[\alpha, \beta]$ within a factor that is polynomial in the input size in digraphs with non-negative edge weights. It is easy to see that as k increases, the problem only becomes harder, with $k = n$ corresponding to the search for a Hamiltonian cycle. Thus, the result follows for any $k \geq 1$. □

5 PRACTICAL ALGORITHMS

5.1 Algorithms for MSIC

5.1.1 Karp's Algorithm. Due to its **NP**-hardness, we will use the *maximum mean-cycle* as an approximate solution to MSIC$[\alpha, \beta]$. We first note that the maximum mean-cycle in a non-negatively weighted graph G is equivalent to the minimum mean-cycle in the graph G' obtained by reversing the sign of the edge weights of G.

The problem of finding the *minimum mean-cycle* (MMC) in a graph with real-valued edge weights is well-studied in the literature and admits efficient polynomial algorithms as shown by Karp [14]. Karp's MMC algorithm runs in $\Theta(nm)$ time and needs $\Theta(n^2)$ space on *any* instance. As noted by Dasdan and Gupta [5], there are other algorithms, with worse theoretical bounds, performing significantly better in practice, such as, Howard's algorithm [11] and Young's algorithm [24]. Dasdan and Gupta [5] have given excellent survey of the different algorithms and their performance in practice.

In this paper we use Karp's MMC algorithm as a heuristic for the MSIC$[\alpha, \beta]$ problem, not only due to its ease of implementation but also it still holds one of the best asymptotic running times.

Lemma 4. *Karp's MMC algorithm [14] provides $O(n)$-approximation for MSIC$[\alpha, \beta]$ in arbitrary graphs with non-negative edge weights.*

Proof. Given a directed graph $G = (V, E)$ with non-negative weights, let C_K denote the cycle with maximum mean weight and let C^* denote the optimal solution to MSIC$[\alpha, \beta]$. Then, by using the fact that $\sum_{e \in C_K} w(e)/|C_K| \geq \sum_{e \in C^*} w(e)/|C^*|$, and that $2 \leq |C| \leq n$ for any cycle C, we obtain

$$\begin{aligned}\frac{F(C^*)}{F(C_K)} &= \frac{\sum_{e \in C^*} w(e)}{\alpha|C^*| + \beta} \cdot \frac{\alpha|C_K| + \beta}{\sum_{e \in C_K} w(e)} \\ &\leq \frac{\sum_{e \in C_K} w(e) \cdot \frac{|C^*|}{|C_K|}}{\alpha|C^*| + \beta} \cdot \frac{\alpha|C_K| + \beta}{\sum_{e \in C_K} w(e)} \\ &= \frac{\alpha + \beta/|C_K|}{\alpha + \beta/|C^*|} \leq \frac{\alpha + \beta/2}{\alpha + \beta/n} \leq n.\end{aligned}$$

□

5.1.2 A variant of Karp's algorithm. Although efficient, a direct application of Karp's algorithm to solve MSIC$[\alpha, \beta]$ disregards the information about the parameters α and β. Thus, we propose a natural extension of Karp's algoritm that incorporates the role of the parameters α and β aligned with the objective function of MSIC$[\alpha, \beta]$. To this end, we modify Karp's algorithm to find the node v that minimizes (on the edge-signs reversed graph G') the following:

$$\min_{v \in V} \max_{1 \leq k \leq n} \frac{D_n(v) - D_k(v)}{\alpha(n - k) + n\beta}. \tag{7}$$

Notice that, as in Karp's characterization, the numerator in (7) mimics the weight of a cycle of length $(n - k)$ found for each $v \in V$, so (7) operates with the objective function of MSIC$[\alpha, \beta]$. Similar to Karp's algorithm, this algorithm runs in $\Theta(nm)$ time and the cycle for the minimizer v can be found by traversing the edge progression $D_n(v)$.

5.2 Algorithms for k-MSIC

The k-MSIC$[\alpha, \beta]$ problem is reminiscent of Steiner cycle problems, thus, one could consider the solutions of related problems, such as maximum mean Steiner cycle (MMSCP), for approximating k-MSIC$[\alpha, \beta]$. However, as we have shown in Section 4, besides being **NP**-hard, both problems cannot be approximated within a ratio that is polynomial in the number of vertices.

Existing algorithms for approximating Steiner cycle problem variants are less well-known, and in most cases these algorithms have strict requirements as we review next.

Steinová [22] proposed a $\frac{3}{2}\log_2(k)$-approximation algorithm for the *mininum* Steiner cycle problem on k terminal vertices in non-negatively weighted graphs in which the edge weights satisfy the triangle inequality. Clearly, this is different from our setting. The instances of k-MSIC$[\alpha, \beta]$ do not satisfy the triangle inequality, and we study a maximization problem.

Salazar-Gonzalez [20] introduced a *minimum* Steiner cycle problem variant with vertex penalties and consider a 0-1 integer linear program examining the Steiner cycle polytope. Besides having a different context, their method is of theoretical interest that doesn't translate into practical algorithms for k-MSIC$[\alpha, \beta]$.

Kanellakis and Papadimitriou [13] propose a *local search* method for directed TSP, extending the Lin-Kernighan heuristic proposed for undirected TSP [18]. We adopt the local search approach proposed by Kanellakis and Papadimitriou [13] for directed TSP and extend their techniques for finding Steiner cycles of interest. We will refer to our local search heuristic for k-MSIC$[\alpha, \beta]$ as Local-SCS.

The local search method by Kanellakis and Papadimitriou [13] starts with a random initial solution then considers the so-called "sequential primary" and "quad" changes. In a sequential primary change, three edges (a, b), (c, d), and (e, f), encountered in this order on the cycle, are removed from the cycle, and the edges (a, d), (c, f) and (e, b) are added. In a quad change, the rewiring consists of removing four edges and reconnecting opposite edges, as shown in Figure 4(b). The neighborhood of each step in their local search consists of a cost-dependent subset, determined by a number of heuristic rules. The search stops when no significant improvements can be made.

When transforming this search from a TSP setting to a Steiner cycle setting, a few adjustments have to be made. Besides the primary and quad change, we consider two new changes in Local-SCS. The *shortcutting change* shortcuts the initial solution into a smaller Steiner cycle. The *extending change* bypasses an edge in a Steiner cycle, by replacing the edge with two new edges. A visualization of all the changes considered by Local-SCS are provided in Figure 4.

Given a set Q of k terminal vertices and an upper bound $l_{max} \geq k$ on the cycle length, Local-SCS finds an initial Steiner cycle of G as follows:

(1) Prune G by only considering each vertice $v \in V$ s.t.

$$\forall q \in Q : \ell(q \rightsquigarrow v) + \ell(v \rightsquigarrow q) \leq l_{max},$$

where $\ell(\cdot)$ denotes the (unweighted) shortest path length. This step can be performed in time $O(k(n + m))$.

(2) Run a randomized depth-first search to find an initial valid Steiner cycle. The search is guided by a heuristic, and each v that has a low total distance towards all query nodes has a higher chance of being explored first, i.e., at any time in the depth-first search, the probability that a vertex v is chosen from the stack is proportional to $1/\sum_{q \in Q} \ell(v \rightsquigarrow q)$.

After Local-SCS finds an initial Steiner cycle, a sequence of changes depicted Figure 4 are applied. When considering a type of change, Local-SCS always selects the one that yields the largest improvement to the objective function (3). Local-SCS first applies a number of extending changes to the initially found cycle until the cycle length is equal to l_{max}. Then, Local-SCS greedily keeps selecting the best change among the sequantial, quad, or shortcutting changes until no improvements can be made. If Local-SCS doesn't return a solution, then no Steiner of length at most l_{max} exists for the given k terminal vertices. The idea is to run this randomized procedure a couple of times (1-5 in the experiments), and pick the best solution.

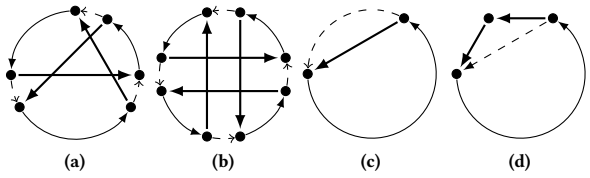

Figure 4: (a) Sequential primary change (b) Quad change (c) Shortcutting change (d) Extending change.

Unlike the method of Kanellakis and Papadimitriou [13], a neighborhood in our local search will consist of *all* the possible changes. For a Steiner cycle of length l_{max}, there are $O(l_{max}^2)$ shortcutting changes, $O(n \cdot l_{max})$ extending changes, $O(l_{max}^3)$ primary changes and $O(l_{max}^4)$ quad changes, which is feasible to evaluate for a reasonable upper bound l_{max}.

6 EXPERIMENTS

The goal of this section is manifold. First, we would like to evaluate the quality of solutions obtained by Karp's MMC algorithm, the variant from Section 5.1.2 and our local Steiner cycle search heuristic Local-SCS. To this end, we conduct experiments on small synthetic datasets and compare the subjective interestingness of the approximate solutions against the optimal solutions that we obtain by exhaustive search in these small instances. Finally, we provide a practical use case: a financial trade dataset between countries. Our Python and Matlab code is publicly available.[5]

6.1 Quality experiments on synthetic datasets

In this section we evaluate the quality of solutions obtained by the algorithms for MSIC$[\alpha, \beta]$, and our local Steiner cycle search heuristic k-MSIC$[\alpha, \beta]$, using various choices of α and β. This requires to exhaustively search for their optimal solutions, by enumerating all the cycles using Johnson's algorithm [12] that runs in $O((n + m)(c + 1))$ time, where c is the total number of cycles in the input graph. To keep the exhaustive search feasible, we perform the quality tests on small instances and generated 200 random Erdős-Rényi graphs with $n = 20$ and edge probability 0.2. Even in such small instances, we found an average of 218,080 cycles per instance, with the maximum number of cycles found in an instance being more than 5 million. We set the weight of each edge to a random integer that is generated uniformly at random from the interval $[1, 10K]$.

We start by evaluating the quality of solutions obtained by Karp's algorithm and its variant for MSIC$[\alpha, \beta]$. We use varying values of α and β obtained by evaluating (2) for $q \in \{0.1, 0.2, 0.3\}$. Figure 5 shows the relative performance w.r.t. the optimal solution for different values of q over 200 random Erdős-Rényi instances,

[5] https://www.dropbox.com/sh/udnimj0uithtxwr/AABvSqaeJNI5-L7TXpU1HNeCa?dl=0. Instructions for reproducibility are provided in the readme.txt file.

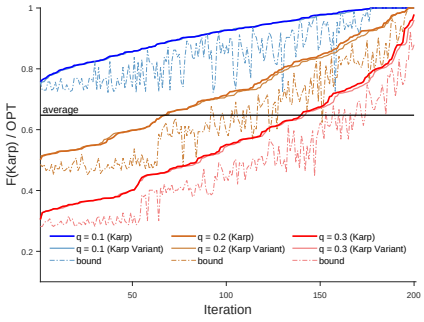

Figure 5: Relative performance of Karp's MMC & Karp's Variant for various q. The dashed lines indicate the theoretical bound provided in Lemma 4.

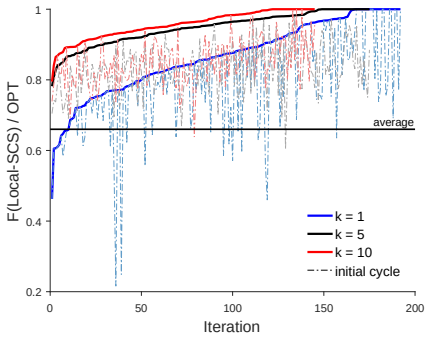

Figure 6: Relative performance of Local-SCS for various k and $q = 0.05$. The dashed lines indicate the best initial solutions before applying changes.

sorted from worst to best performing. In order to have a baseline, we compute the average interestingness over all possible cycles that were encountered in the 200 instances. The influence of the parameter q is clearly visible. For increasing q, the performance drops drastically as expected, since, the optimal cycle corresponds more to the longest weighted cycle, while Karp's algorithm provides the cycle with maximum mean weight. Interestingly, the variant from Section 5.1.2 performs slightly worse overall than Karp's algorithm. However, we report that in a small number of instances it performed significantly better than Karp's algorithm although this trend didn't generalize. As a guideline, we advise to set q to be not larger than the density of the network (which is 0.2 in this case).

Next we evaluate the quality of solutions obtained by Local-SCS for k-MSIC$[\alpha, \beta]$. We set $q = 0.05$ and randomly pick k terminal vertices, for $k \in \{1, 5, 10\}$. We set no upper bound on the maximum cycle length, i.e., $l_{max} = 20$, run the algorithm 5 times, and pick the best solution. Relative performance is shown in Figure 6. Instances in the x-axis are again sorted from worse to best performing. The dashed lines indicate the best value of an initial Steiner cycle that was found in the 5 tries, clearly showing that the sequence of changes proposed in Section 5.2 improve the score by a good amount. We also observed that Local-SCS didn't find any Steiner cycle in 55 out of 200 instances for $k = 10$, while this number was 25 for $k = 5$ and 8 for $k = 1$. The increase in the performance for larger k is mainly due to the fact that there are more possible local changes available to perform on an initially found cycle for higher k, provided that a Steiner cycle of length at most l_{max} exists.

We analyze the running time of Local-SCS[6] in two different settings, see Figure 7. First, we generate Erdős graphs of size $n = 20$ with edge probability 0.2, set no bound on l_{max}, and let the query size k vary. For each k, we generate 50 graphs and repeat the algorithm one time. As expected, for fixed n and m, the running time is linear in k. Second, we set $k = 3$, $l_{max} = 10$ and let the graph size n vary. Again for each n, we generate 50 instances. As expected, there is a polynomial dependence on the graph size n; doubling the graph size n roughly leads to a quadrupling in running time.

[6]Karps's MMC and Karp's Variant *always* run in $\Theta(nm)$ time, hence are not tested.

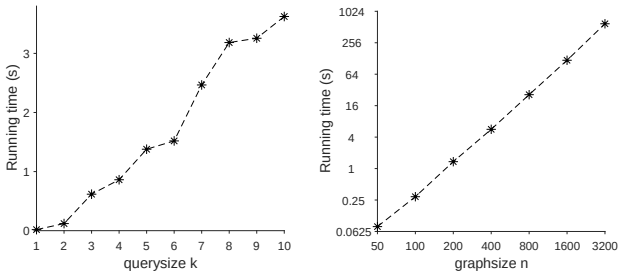

Figure 7: Running times of Local-SCS on Erdős graphs: on the left for varying query size k, on the right for varying graph size n.

6.2 Practical use case

To see the influence of a prior belief model on the resulting cycles, we look at the trading volumes between countries in 2018. This dataset is publicly available, and its basic statistics are summarized in Table 1.

Table 1: Network statistics.

Dataset	|V|	|E|	Edge Weights
Trade[7]	221	1 957	Country top 5 import & export in 2018

We set $l_{max} = 6$, $q = 0.01$, and used 10 iterations of Local-SCS. First, we fit a geometric model with the weighted in- and outdegree of each node as a prior. Figure 8a shows the most interesting cycle in the graph: a 2-cycle between the U.S. and the Dominican Republic. The % outside the circle denotes the weight of an edge (u, v), relative to the total export of u. The % inside the circle the denotes the weight relative to the total import of v. As discussed in Section 2.4, these edges are indeed very interesting: the Dominican Republic is extremely economically-dependent on the U.S. in terms of import and export. However, the converse it not true. Figure 8b shows the most interesting cycle when we take the bilateral trade

[7]https://wits.worldbank.org/

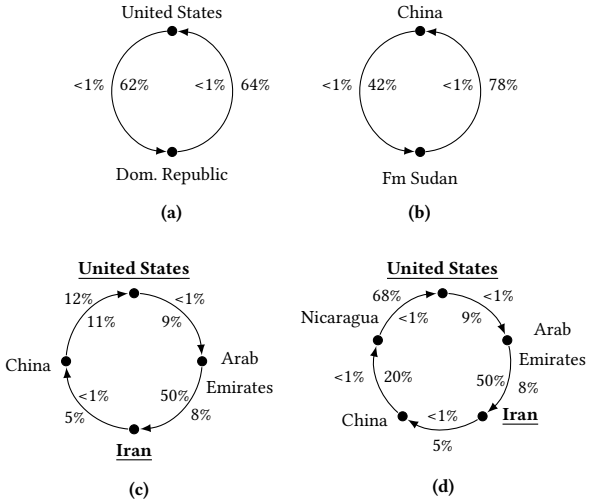

Figure 8: The most interesting cycles: (a) with a prior on weighted in- and out-degree of each country (b) with a prior on the trading volume between US and Dom. Rep. (c) with Iran and US as query nodes, with a prior on weighted in- and out-degrees of each country (d) Iran and US as query nodes, with a prior on trading volume between US and China.

agreement between the U.S. and Dominican Republic into account as a prior belief. Since these edges are now more expected, they become less interesting. The new most interesting cycle is another 2-cycle, between China and Sudan. Again, a small country that is economically-dependent on a bigger country. Figure 8c shows the result when we query both Iran and the U.S., two countries not expected to be in a direct trade relationship because of the U.S. trade embargo on Iran. This cycle now contains China as an export country for Iran and China linking back to the U.S. Figure 8d shows the result when we take the trade relationships between the U.S. and China into account as well. The direct edge is now expected, and the resulting heuristic takes this into account by placing an intermediate country in between, Nicaragua. Nicaragua heavily depends on China for its import, and the U.S. for its export, thus making these connections interesting.

7 RELATED WORK

Discovering cyclic patterns in graphs has not received much attention in the data-mining community. Giscard et al. [10] evaluate the balance of a signed social network by finding simple cycles. Kumar and Calders [15] propose an algorithm for enumerating all simple cycles in a directed temporal network, by extending Johnson's algorithm [12] to a temporal setting. Our aim in this paper is to discover the most interesting cycles with respect to a *subjective interestingness* measure, a concept introduced by Silberschatz and Tuzhilin [21], and later extended by De Bie [6]. Building on that framework, van Leeuwen et al. [23] studied the problem of subjectively interesting subgraph pattern mining and Adriaens et al. [1] studied subjectively interesting Steiner trees.

8 CONCLUSIONS & FUTURE WORK

In this paper, we introduce the problem of discovering interesting cycles in digraphs. We formally define the problems of finding the maximum subjectively interesting cycles and Steiner cycles. We provide an extensive computational complexity analysis for both problems and propose a number of efficient heuristics. We experimentally verify the effectiveness of our methods and provide a real-world use case. Our work opens interesting directions for future research. First, it is worth to consider the usefulness of a non-simple cycle (a *tour*) as a data-mining pattern. Second, we aim to extend our results for discovering cycles in undirected graphs, which is a non-trivial extension of the directed case.

REFERENCES

[1] F. Adriaens, J. Lijffijt, and T. De Bie. 2017. Subjectively interesting connecting trees. In *ECML PKDD 2017*, Vol. 10535. Springer International Publishing, 53–69.
[2] A. Björklund, T. Husfeldt, and S. Khanna. 2004. Approximating Longest Directed Paths and Cycles. In *Automata, Languages and Programming*. Springer Berlin.
[3] A. F. Colladon and E. Remondi. 2017. Using social network analysis to prevent money laundering. *Expert Systems with Applications* 67 (2017), 49–58.
[4] P. Crescenzi. 1997. A Short Guide to Approximation Preserving Reductions. In *IEEE Conference on Computational Complexity*.
[5] A. Dasdan and R. K. Gupta. 2006. Faster Maximum and Minimum Mean Cycle Algorithms for System-performance Analysis. *TCADICS* 17, 10 (Nov. 2006).
[6] T. De Bie. 2011. An information theoretic framework for data mining. In *Proc. KDD*. 564–572.
[7] T. De Bie. 2011. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *DMKD* 23 (2011), 407–446.
[8] J. A. Dunne, R. J. Williams, and N. D. Martinez. 2002. Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology letters* (2002).
[9] H. N. Gabow and S. Nie. 2004. Finding a Long Directed Cycle. In *ACM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics.
[10] P.L. Giscard, P. Rochet, and R. C. Wilson. 2017. Evaluating balance on social networks from their simple cycles. *J. Complex Networks* 5 (2017).
[11] R. A. Howard. 1960. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA.
[12] D. Johnson. 1975. Finding All the Elementary Circuits of a Directed Graph. *SIAM J. Comput.* 4, 1 (1975), 77–84.
[13] P. C. Kanellakis and C. H. Papadimitriou. 1980. Local Search for the Asymmetric Traveling Salesman Problem. *Operations Research* 28, 5 (1980), 1086–1099.
[14] R. M. Karp. 1978. A characterization of the minimum cycle mean in a digraph. *Discrete Mathematics* 23, 3 (1978), 309 – 311.
[15] R. Kumar and T. Calders. 2017. Finding simple temporal cycles in an interaction network. In *TD-LSG@PKDD/ECML*.
[16] Y.K. Kwon and K. H. Cho. 2007. Analysis of feedback loops and robustness in network evolution based on Boolean models. *BMC bioinformatics* (2007).
[17] M. van Leeuwen, T. De Bie, E. Spyropoulou, and C. Mesnage. 2016. Subjective interestingness of subgraph patterns. *Mach. Learn.* 105, 1 (2016), 41–75.
[18] S. Lin and B. W. Kernighan. 1973. An Effective Heuristic Algorithm for the Traveling-Salesman Problem. *Oper. Res.* 21, 2 (April 1973), 498–516.
[19] Yehoshua Perl and Yossi Shiloach. 1978. Finding Two Disjoint Paths Between Two Pairs of Vertices in a Graph. *J. ACM* 25 (1978), 1–9.
[20] Juan-Jose Salazar-Gonzalez. 2003. The Steiner cycle polytope. *European Journal of Operational Research* 147, 3 (2003), 671 – 679.
[21] A. Silberschatz and A. Tuzhilin. 1996. On subjective measures of interestingness in knowledge discovery. In *Proc. KDD*. 275–281.
[22] Monika Steinová. 2010. Approximability of the minimum Steiner cycle problem. *Computing and Informatics* 29 (2010), 1349–1357. Issue 6+.
[23] M. van Leeuwen, T. De Bie, E. Spyropoulou, and C. Mesnage. 2016. Subjective interestingness of subgraph patterns. *Machine Learning* 105, 1 (2016), 41–75.
[24] Neal E. Young, Robert E. Tarjan, and James B. Orlin. 1991. Faster Parametric Shortest Path and Minimum Balance Algorithms. *Networks* 21 (1991), 205–221.