Nr rejestracyjny: 2020/39/D/ST6/00419; Kierownik projektu: dr inż. Krzysztof Szymon Turowski

Charakteryzacja treści informacyjnej struktur grafowych

Krzysztof Turowski

Współcześnie grafy są szeroko stosowane do opisu różnych struktur: sieci społecznościowych, interakcji protein, a nawet zależności funkcjonalnych w mózgu. Sieci te bardzo często mają wiele milionów wierzchołków o najróżniejszych interakcjach. Powstaje naturalne pytanie o możliwość efektywnego przechowywania, dostępu i przetwarzania danych tego rodzaju.

Oczywiście w przypadku rzeczywistych sieci nie mamy bezpośredniego dostępu do wiedzy o tym, zgodnie z jakim losowym modelem wyewoluowała. Możemy jednak posłużyć się szerszą wiedzą np. biologiczną o konkretnych badanych procesach oraz wnioskowaniem z rzeczywistych sieci, jakie są możliwe zakresy parametrów, które dawałyby jak największą szansę ich wygenerowania. Przykładowo, dla grafów opisujących interakcje protein wśród naukowców można zaobserwować szeroką zgodę, że u podstaw leżą mechanizmy duplikacji i mutacji genów. Te cechy dobrze ujmuje duplikacyjny model grafów losowych, polegający na rozbudowywaniu grafu przez wybór losowego wierzchołka, stworzenie jego kopii, a następnie dodawanie i usuwanie krawędzi zgodnie z pewnymi ustalonymi regułami. Jednym z celów naszych badań jest:

(1) weryfikacja tych hipotez w odniesieniu do różnych modeli duplikacyjnych w zestawieniu z innymi popularnymi w literaturze modelami.

Klasyczna teoria informacji wprowadziła entropię jako naturalną definicję treści informacyjnej pewnego rozkładu prawdopodobieństwa nad abstrakcyjnym zbiorem obiektów. Można zastosować to pojęcie również do grafów i pytać o entropię grafu z etykietami, czyli ile miejsca potrzebujemy, by zapisać graf z etykietami. Można pytać również o entropię *struktury* grafu, a więc ile pamięci możemy zyskać, gdy nie musimy przechowywać informacji o wierzchołkach, a tylko o relacjach między nimi. Znajomość obu parametrów jest pomocna do oceny jakości proponowanych algorytmów kompresji, a nawet może być pomocna w znalezieniu algorytmu optymalnego. Dlatego kolejnym celem badań będzie:

(2) oszacowanie entropii oraz entropii strukturalnej grafów losowych dla modeli duplikacyjnych oraz opracowanie szybkich algorytmów kompresji o ustalonych gwarancjach aproksymacji.

Pytanie o strukturę w naturalny sposób jest związane z poszukiwaniem symetrii oraz regularności struktury w grafie. Pierwsza cecha grafu jest opisywana przez pojęcie grupy automorfizmów, czyli takiej zamiany etykiet wierzchołków, że graf pozostaje bez zmian: jeśli istniała krawędź między wierzchołkami o etykietach a i b przed zamianą, to również po zamianie a i b są połączone. Drugą własność często można wyrazić przez takie parametry grafowe jak rozkład stopni wierzchołków czy też rozkład występowania prostych podstruktur np. trójkątów. Przykładowo, jeśli w grafie bardzo często wiele wierzchołków ma dokładnie jednego i tego samego sąsiada, to intuicyjnie widać, że powinien istnieć zwięzły zapis reprezentujący taką podstrukturę. Punktem wyjścia do badań powyższych problemów będą następujące zagadnienia:

(3) badanie rozkładów częstości występowania małych struktur w grafie losowym według modelu duplikacyjnego,

(4) analiza rozkładów stopni w grafach losowych wygenerowanych z modelu duplikacyjnego,

(5) szacowanie rozkładu symetrii w grafach losowych dla modeli duplikacyjnych.