GGDs: Graph Generating Dependencies

Larissa C. Shimomura
l.capobianco.shimomura@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

George Fletcher
g.h.l.fletcher@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Nikolay Yakovets
hush@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Constraints play a key role in data management research, e.g., in the study of data quality, data integration and exchange, and query optimization. As graph-structured data sets are commonly used in a diverse number of applications, the study of graph dependencies is also of increasing interest. Recently, different classes of dependencies for graphs have been proposed such as Graph Functional Dependencies (GFDs [3]), Graph Entity Dependencies (GEDs [2]), and Graph Differential Dependencies (GDDs [4]). However, these dependencies focus on generalizing functional dependencies (i.e., variations of *equality*-generating dependencies) and cannot capture *tuple*-generating dependencies (TGDs) for graph data [1].

We introduce a new class of graph dependencies called Graph Generating Dependencies (GGDs)[5] which supports TGDs for property graphs. A GGD expresses a constraint between two (possibly) different graph patterns enforcing relationships between property values and topological structure.

The main differences of our proposed GGDs compared to previous works are the use of differential constraints (on both source and target side), edges are treated as first-class citizens in the graph patterns (in alignment with the property graph model), and the ability to entail the generation of new vertices and edges. With these new features of the GGDs, we can encode relations between two graph patterns as well as the (dis)similarity between its vertices and edges properties values.

A graph generating dependency (GGDs) is defined as

$$Q_s[\overline{x}], \phi_s \rightarrow Q_t[\overline{x}, \overline{y}], \phi_t$$

where: (1) $Q_s[\overline{x}]$ and $Q_t[\overline{x}, \overline{y}]$ are graph patterns, called **source** graph pattern and **target** graph pattern, respectively; (2) ϕ_s is a set of differential constraints defined over the variables $\overline{x}$ (variables of the graph pattern Q_s); and ϕ_t is a set of differential constraints defined over the variables $\overline{x} \cup \overline{y}$, in which $\overline{x}$ are the variables of the source graph pattern Q_s and $\overline{y}$ are any additional variables of the target graph pattern Q_t.

A differential constraint in ϕ_s on $[\overline{x}]$ (resp., in ϕ_t on $[\overline{x}, \overline{y}]$) is a constraint of one of the following forms [4, 6]: (1)$\delta_A(x.A, c) \leq t_A$; (2) $\delta_{A_1A_2}(x.A_1, x'.A_2) \leq t_{A_1A_2}$ and (3)$x = x'$ or $x \neq x'$ where $x, x' \in \overline{x}$ (resp. $\in \overline{x} \cup \overline{y}$) for $Q_s[\overline{x}]$ (resp. for $Q_t[\overline{x}, \overline{y}]$), δ_A is a user defined similarity function for the property A (resp. $\delta_{A_1A_2}$ us a user defined function for the properties A_1, A_2) and $x.A$ is the property value of variable x on A, c is a constant of the domain of property A and t_A, $t_{A_1A_2}$ are pre-defined thresholds.

A GGD $\sigma = Q_s[\overline{x}], \phi_s \rightarrow Q_t[\overline{x}, \overline{y}], \phi_t$ holds in a graph G, denoted as $G \models \sigma$, if and only if for every homomorphic graph pattern match $h_s[\overline{x}]$ of the source graph pattern $Q_s[\overline{x}]$ in G satisfying the set of constraints ϕ_s, there exists a homomorphic match $h_t[\overline{x}, \overline{y}]$ of the graph pattern $Q_t[\overline{x}, \overline{y}]$ in G satisfying ϕ_t such that for each x in $\overline{x}$ it holds that $h_s(x) = h_t(x)$. In case a GGD is not satisfied, we typically fix this by *generating* new vertices/edges in G. See an example of a Graph Generating Dependencies on Figure 1.

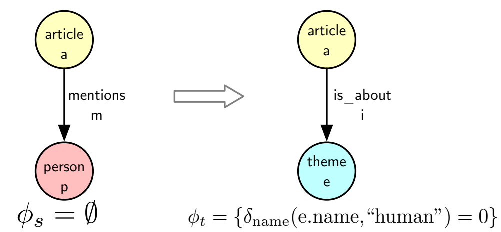

Figure 1: Example of GGD: If an article a mentions a person p then the same article a should have an edge labelled “is_about" to a node theme in which the distance between the theme name and the string “human" is zero.

Based on the semantics of the GGDs, we develop an algorithm for solving the validation problem for GGDs. The validation problem consists of checking if, given a finite set Σ of GGDs and graph G, $G \models \Sigma$.

GGDs can be used for different applications, one of the applications is Entity Resolution (ER) . As mentioned beforehand, the main novelty of the GGDs is the generation of new vertices or edges in case a GGD is not validated. Given this feature, it is possible to rewrite ER matching rules or conditions as GGDs.

Towards entity resolution, we can define the source graph patterns as several disjoint patterns from (possibly) different graph sources and use the target graph pattern specifications as the representation of the deduplicated graphs. Thus, using this approach, we can also encode more information than just vertex-to-vertex, or row-to-row in relational databases, as we consider all the information in a defined graph pattern.

Acknowledgments. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825041.

REFERENCES

[1] Wenfei Fan. 2019. Dependencies for Graphs: Challenges and Opportunities. *J. Data and Information Quality* 11, 2 (2019), 5:1–5:12.

[2] Wenfei Fan and Ping Lu. 2019. Dependencies for Graphs. *ACM Trans. Database Syst.* 44, 2, Article 5 (Feb. 2019), 40 pages.

[3] Wenfei Fan, Yinghui Wu, and Jingbo Xu. 2016. Functional Dependencies for Graphs. In *SIGMOD*. 1843–1857.

[4] Selasi Kwashie, Lin Liu, Jixue Liu, Markus Stumptner, Jiuyong Li, and Lujing Yang. 2019. Certus: An Effective Entity Resolution Approach with Graph Differential Dependencies (GDDs). *Proc. VLDB Endow.* 12, 6 (Feb. 2019), 653–666.

[5] Larissa C. Shimomura, George Fletcher, and Nikolay Yakovets. 2020. GGDs: Graph Generating Dependencies. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. 2217–2220.

[6] Shaoxu Song and Lei Chen. 2011. Differential Dependencies: Reasoning and Discovery. *ACM Trans. Database Syst.* 36, 3, Article 16 (Aug. 2011), 41 pages.