Chapter 16
InfoNetOLAP: OLAP and Mining of Information Networks

Chen Chen, Feida Zhu, Xifeng Yan, Jiawei Han, Philip Yu, and Raghu Ramakrishnan

Abstract Databases and data warehouse systems have been evolving from handling normalized spreadsheets stored in relational databases to managing and analyzing diverse application-oriented data with complex interconnecting structures. Responding to this emerging trend, information networks have been growing rapidly and showing their critical importance in many applications, such as the analysis of XML, social networks, Web, biological data, multimedia data, and spatiotemporal data. Can we extend useful functions of databases and data warehouse systems to handle network structured data? In particular, OLAP (On-Line Analytical Processing) has been a popular tool for fast and user-friendly *multi-dimensional* analysis of data warehouses. *Can we OLAP information networks and perform mining tasks on top of that?* Unfortunately, to our best knowledge, there are no OLAP tools available that can interactively view and analyze network structured data from different perspectives and with multiple granularities. In this chapter, we argue that it is critically important to OLAP such information network data and propose a novel *InfoNetOLAP* framework. According to this framework, given an information network data set with its nodes and edges associated with respective attributes, a multi-dimensional model can be built to enable efficient on-line analytical processing so that *any portions* of the information networks can be *generalized/specialized* dynamically, offering multiple, versatile views of the data set. The contributions of this work are threefold. First, starting from basic definitions, i.e., what are *dimensions* and *measures* in the InfoNetOLAP scenario, we develop a conceptual framework for data cubes constructed on the information networks. We also look into different semantics of OLAP operations and classify the framework into two major subcases: *informational OLAP* and *topological OLAP*. Second, we show how an information network cube can be materialized by calculating a special kind of measure called *aggregated graph* and how to implement it efficiently. This includes both full materialization and partial materialization where constraints are enforced to obtain an *iceberg cube*. As we can see, due to the increased structural complexity

C. Chen (✉)
University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: cchen37@uiuc.edu

P.S. Yu, et al. (eds.), *Link Mining: Models, Algorithms, and Applications*,
DOI 10.1007/978-1-4419-6515-8_16, © Springer Science+Business Media, LLC 2010

of data, aggregated graphs that depend on the underlying "graph" properties of the information networks are much harder to compute than their traditional OLAP counterparts. Third, to provide more flexible, interesting, and insightful OLAP of information networks, we further propose a *discovery-driven* multi-dimensional analysis model to ensure that OLAP is performed in an intelligent manner, guided by expert rules and knowledge discovery processes. We outline such a framework and discuss some challenging research issues for discovery-driven InfoNetOLAP.

16.1 Introduction

OLAP (On-Line Analytical Processing) [2, 7, 11, 12, 31] is an important notion in data analysis. Given the underlying data, a cube can be constructed to provide a *multi-dimensional* and *multi-level* view, which allows for effective analysis of the data from different perspectives and with multiple granularities. The key operations in an OLAP framework are slice/dice and roll-up/drill-down, with slice/dice focusing on a particular aspect of the data, roll-up performing generalization if users only want to see a concise overview, and drill-down performing specialization if more details are needed.

In a traditional data cube, a data record is associated with a set of dimensional values, whereas different records are viewed as *mutually independent*. Multiple records can be summarized by the definition of corresponding aggregate measures such as COUNT, SUM, and AVERAGE. Moreover, if a concept hierarchy is associated with each attribute, multi-level summaries can also be achieved. Users can navigate through different dimensions and multiple hierarchies via roll-up, drill-down, and slice/dice operations. However, in recent years, more and more data sources beyond conventional spreadsheets have come into being, such as chemical compounds or protein networks (chem/bio-informatics), 2D/3D objects (pattern recognition), circuits (computer-aided design), XML (data with loose schema), and Web (human/computer networks), where not only individual entities but also the *interacting relationships* among them are important and interesting. This demands a new generation of tools that can manage and analyze such data.

Given their great expressive power, information networks have been widely used for modeling a lot of data sets that contain structure information. With the tremendous amount of information network data accumulated in all above applications, the same need to deploy analysis from different perspectives and with multiple granularities exists. To this extent, our main task in this chapter is to develop an *InfoNetOLAP framework*, which presents a multi-dimensional and multi-level view over information networks.

In order to illustrate what we mean by "InfoNetOLAP" and how the OLAP glossary is interpreted with regard to this new scenario, let us start from a few examples.

Example 1 (Collaboration Patterns) There are a set of authors working in a given field: For any two persons, if they coauthor w papers in a conference, e.g.,

SIGMOD 2004, then a link is added between them, which has a collaboration frequency attribute that is weighted as w. For every conference in every year, we may have a coauthor network describing the collaboration patterns among researchers; each of them can be viewed as a snapshot of the overall coauthor network in a bigger context.

It is interesting to analyze the aforementioned network data set in an OLAP manner. First, one may want to check the collaboration patterns for a group of conferences, say, all DB conferences in 2004 (including SIGMOD, VLDB, ICDE, etc.) or all SIGMOD conferences since the year it was introduced. In the language of data cube, with a *venue* dimension and a *time* dimension, one may choose to obtain the *(db-conf, 2004)* cell and the *(sigmod, all-years)* cell, where the *venue* and *time* dimensions have been generalized to *db-conf* and *all-years*, respectively. Second, for the subset of snapshots within each cell, one can summarize them by computing a measure as we did in traditional OLAP. In the InfoNetOLAP context, this gives rise to an *aggregated graph*. For example, a summary network displaying total collaboration frequencies can be achieved by overlaying all snapshots together and summing up the respective edge weights, so that each link now indicates two persons' collaboration activities in the DB conferences of 2004 or during the whole history of SIGMOD. ■

The above example is simple because the measure is calculated by a simple sum over individual pieces of information. A more complex case is presented next.

Example 2 (Maximum Flow) Consider a set of cities connected by transportation networks. In general, there are many ways to go from one city A to another city B, e.g., by bus, by train, by air, by water, and each is operated by multiple companies. For example, we can assume that the capacity of company x's air service from A to B is c^x_{AB}, i.e., company x can transport at most c^x_{AB} units of cargo from A to B using the planes it owns. Finally, we get a snapshot of capacity network for every transportation means of every company.

Now, consider the transporting capability from a source city S to a destination city T; it is interesting to see how this value can be achieved by sending flows of cargo through different paths if (1) we only want to go by air or (2) we only want to choose services operated by company x. In the OLAP language, with a *transportation-type* dimension and a *company* dimension, the above situations correspond to the *(air, all-companies)* cell and the *(all-types, company x)* cell, while the *measure* computed for a cell c is a graph displaying how the maximum flow can be configured, which has considered all transportation means and operating companies associated with c. Unlike Example 1, computing the aggregated graph of maximum flow is now a much harder task; also, the semantics associated with un-aggregated network snapshots and aggregated graphs are different: The former shows capacities on its edges, whereas the latter shows transmitted flows, which by definition must be smaller. ■

Example 3 (Collaboration Patterns, Revisited) Usually, the whole coauthor network could be too big for the users to comprehend, and thus it is desirable to look at a

more compressed view. For example, one may like to see the collaboration activities organized by the authors' associated affiliations, which requires the network to be generalized one step up, i.e., merging all persons in the same institution as one node and constructing a new summary graph at the institution level. In this "generalized network", for example, an edge between Stanford and University of Wisconsin will aggregate all collaboration frequencies incurred between Stanford authors and Wisconsin authors. Similar to Examples 1 and 2, an aggregated graph (i.e., the generalized network defined above) is taken as the OLAP measure. However, the difference here is that a roll-up from the individual level to the institution level is achieved by consolidating multiple nodes into one, which shrinks the original network. Compared to this, the graph in Examples 1 and 2 is not collapsed because we are always examining the relationships among the same set of objects; it poses minimum changes with regard to network topology upon generalization. ■

The above examples demonstrate that OLAP provides a powerful primitive to examine information networks. In this chapter, we will give a systematic study on InfoNetOLAP, which is more general than the traditional OLAP: In addition to individual entities, the mutual interactions among them are also considered in the analytical process. Our major contributions are summarized below.

1. *On conceptual modeling*, an *InfoNetOLAP framework* is developed, which defines *dimensions* and *measures* in the information network context, as well as the concept of *multi-dimensional* and *multi-level* analysis over network structured data. We distinguish different semantics of OLAP operations and categorize them into two major subcases: *informational OLAP* (as shown in Examples 1 and 2) and *topological OLAP* (as shown in Example 3). It is necessary since these two kinds of OLAP demonstrate substantial differences with regard to the construction of data cubes over information networks.
2. *On efficient implementation*, the computation of *aggregated graphs* as InfoNetOLAP measures is examined. Due to the increased structural complexity of data, calculating certain measures that are closely tied with the "graph" properties of the information networks, e.g., maximum flow, centrality, poses greater challenges than their traditional OLAP counterparts, such as COUNT, SUM, and AVERAGE. We investigate this issue, categorize measures based on the difficulty to compute them in the OLAP context, and suggest a few measure properties that might help further optimize processings. Both full materialization and partial materialization (where constraints are enforced to obtain an *iceberg cube*) are discussed.
3. *On utilizing the framework for knowledge discovery*, we propose *discovery-driven InfoNetOLAP*, so that interesting patterns and knowledge can be effectively discovered. After presenting the general ideas, we outline a list of guiding principles and discuss some associated research problems. Being part of an ongoing project, it opens a promising path that could potentially lead to more flexible and insightful OLAP of information networks, which we shall explore further in future works.

The remainder of this chapter is organized as follows. In Section 16.2, we formally introduce the InfoNetOLAP framework. Section 16.3 discusses the general hardness to compute aggregated graphs as InfoNetOLAP measures, which categorizes them into three classes. Section 16.4 looks into some properties of the measures and proposes a few computational optimizations. Constraints and partial materialization are studied in Section 16.5. Discovery-driven InfoNetOLAP is covered in Section 16.6. We report experiment results and related work in Sections 18.7 and 16.8, respectively. Section 16.9 concludes this study.

16.2 An InfoNetOLAP Framework

In this section, we present the general framework of InfoNetOLAP.

Definition 1 (Information Network Model) We model the data examined by InfoNetOLAP as *a collection of network snapshots* $\mathcal{G} = \{\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_N\}$, where each snapshot $\mathbb{G}_i = (I_{1,i}, I_{2,i}, \ldots, I_{k,i}; G_i)$ in which $I_{1,i}, I_{2,i}, \ldots, I_{k,i}$ are k *informational attributes* describing the snapshot as a whole and $G_i = (V_i, E_i)$ is a *graph*. There are also *node attributes* attached with any $v \in V_i$ and *edge attributes* attached with any $e \in E_i$. Note that since $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_N$ only represent different observations, $V_1, V_2, \ldots, V_N$ actually correspond to the same set of objects in real applications.

For instance, with regard to the coauthor network described in the introduction, *venue* and *time* are two informational attributes that mark the status of individual snapshots, e.g., SIGMOD 2004 and ICDE 2005, *authorID* is a node attribute indicating the identification of each node, and *collaboration frequency* is an edge attribute reflecting the connection strength of each edge.

Dimension and *measure* are two concepts that lay the foundation of OLAP and cubes. As their names imply, first, dimensions are used to construct a cuboid lattice and partition the data into different cells, which act as the basis for multi-dimensional and multi-level analysis; second, measures are calculated to aggregate the data covered, which deliver a summarized view of it. Below, we are going to formally re-define these two concepts concerning the InfoNetOLAP scenario.

Let us examine dimensions at first. Actually, there are two types of dimensions in InfoNetOLAP. The first one, as exemplified by Example 1, utilizes informational attributes attached at the whole snapshot level. Suppose the following concept hierarchies are associated with *venue* and *time*:

- *venue*: *conference* $\rightarrow$ *area* $\rightarrow$ *all*,
- *time*: *year* $\rightarrow$ *decade* $\rightarrow$ *all*;

the role of these two dimensions is to organize snapshots into groups based on different perspectives, e.g., *(db-conf, 2004)* and *(sigmod, all-years)*, where each of these groups corresponds to a "cell" in the OLAP terminology. They control what snapshots are to be looked at, without touching the inside of any single snapshot.

Definition 2 (Informational Dimensions) With regard to the information network model presented in Definition 1, the set of informational attributes $\{I_1, I_2, \ldots, I_k\}$ are called the *informational dimensions* of InfoNetOLAP, or *Info-Dims* in short.

The second type of dimensions is provided to operate on nodes and edges within individual networks. Take Example 3 for instance; suppose the following concept hierarchy

- *authorID*: *individual* → *department* → *institution* → *all*

is associated with the node attribute *authorID*, then it can be used to group authors from the same institution into a "generalized" node, and a new network thus formed will depict interactions among these groups as a whole, which summarizes the original network and hides specific details.

Definition 3 (Topological Dimensions) The set of dimensions coming from the attributes of topological elements (i.e., nodes and edges of G_i), $\{T_1, T_2, \ldots, T_l\}$, are called the *topological dimensions* of InfoNetOLAP, or *Topo-Dims* in short.

The OLAP semantics accomplished through Info-Dims and Topo-Dims are rather different, and in the following we shall refer to them as *informational OLAP* (abbr. *I-OLAP*) and *topological OLAP* (abbr. *T-OLAP*), respectively.

For roll-up in *I-OLAP*, the characterizing feature is that snapshots are just different observations of the same underlying network, and thus when they are all grouped into one cell in the cube, it is like *overlaying* multiple pieces of information, *without changing* the objects whose interactions are being looked at.

For roll-up in *T-OLAP*, we are no longer grouping snapshots, and the reorganization switches to happen inside individual networks. Here, *merging* is performed internally which "zooms out" the user's focus to a "generalized" set of objects, and a new information network formed by such *shrinking* might greatly alter the original network's topological structure.

Now we move on to measures. Remember that, in traditional OLAP, a measure is calculated by aggregating all the data tuples whose dimensions are of the same values (based on concept hierarchies, such values could range from the finest ungeneralized ones to "all/*", which form a multi-level cuboid lattice); casting this to our scenario here:

First, in InfoNetOLAP, the aggregation of graphs should also take the form of a graph, i.e., an *aggregated graph*. In this sense, graph can be viewed as a special existence, which plays a dual role: as a data source and as an aggregated measure. Of course, other measures that are not graphs, such as node count, average degree, diameter, can also be calculated; however, we do not explicitly include such non-graph measures in our model, but instead treat them as derived from corresponding graph measures.

Second, due to the different semantics of I-OLAP and T-OLAP, aggregating data with identical Info-Dim values groups information among the snapshots, whereas aggregating data with identical Topo-Dim values groups topological elements inside individual networks. As a result, we will give a separate measure definition for each case below.

Definition 4 (I-Aggregated Graph) With regard to Info-Dims $\{I_1, I_2, \ldots, I_k\}$, the *I-aggregated graph* M^I is an *attributed graph* that can be computed based on a set of network snapshots $\mathcal{G}' = \{\mathbb{G}_{i_1}, \mathbb{G}_{i_2}, \ldots, \mathbb{G}_{i_{N'}}\}$ whose Info-Dims are of identical values; it satisfies the following: (1) the nodes of M^I are as same as any snapshot in $\mathcal{G}'$, and (2) the node/edge attributes attached to M^I are calculated as *aggregate functions* of the node/edge attributes attached to $G_{i_1}, G_{i_2}, \ldots, G_{i_{N'}}$.

The graph in Fig. 16.1 that describes collaboration frequencies among individual authors for a particular group of conferences during a particular period of time is an instance of I-aggregated graph, and the interpretation of classic OLAP operations with regard to InfoNet I-OLAP is summarized as follows:

- Roll-up: Overlay multiple snapshots to form a higher-level summary via I-aggregated graph.
- Drill-down: Return to the set of lower-level snapshots from the higher-level overlaid (aggregated) graph.
- Slice/dice: Select a subset of qualifying snapshots based on Info-Dims.

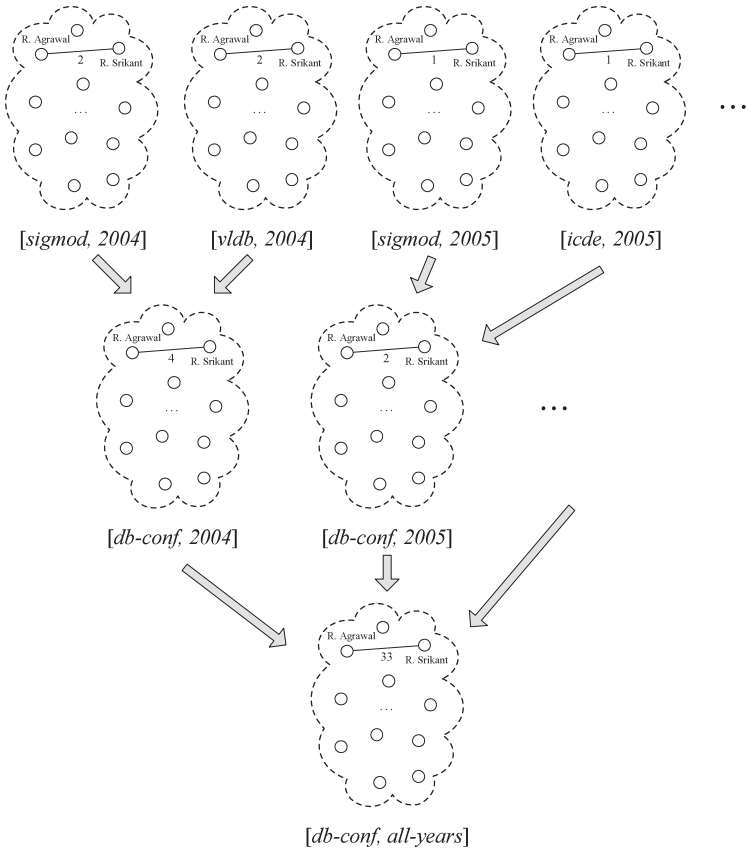

Fig. 16.1 The OLAP scenario for Example 1

Definition 5 (T-Aggregated Graph) With regard to Topo-Dims $\{T_1, T_2 \ldots, T_l\}$, the *T-aggregated graph* M^T is an *attributed graph* that can be computed based on an individual network G_i; it satisfies the following: (1) the nodes in G_i that have identical values on their Topo-Dims are grouped, whereas each group corresponds to a node of M^T, and (2) the attributes attached to M^T are calculated as *aggregate functions* of the attributes attached to G_i.

The graph in Fig. 16.2 that describes collaboration frequencies among institutions is an instance of T-aggregated graph, and the interpretation of classic OLAP operations with regard to InfoNet T-OLAP is summarized as follows:

- Roll-up: Shrink the network topology and obtain a T-aggregated graph that displays compressed views. Topological elements (i.e., nodes and/or edges) are merged and replaced by corresponding higher-level ones during the process.
- Drill-down: A reverse operation of roll-up.
- Slice/dice: Select a subgraph of the network based on Topo-Dims.

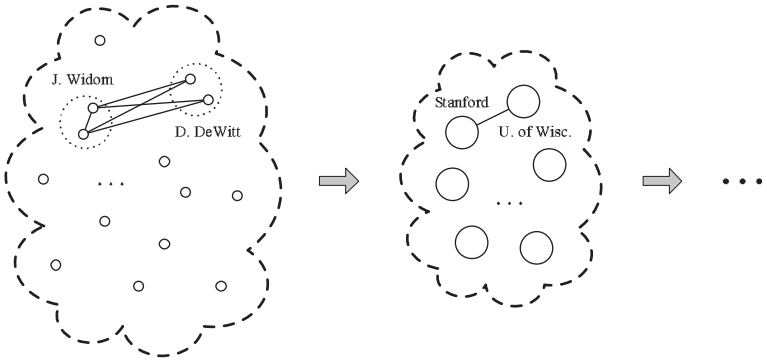

Fig. 16.2 The OLAP scenario for Example 3

16.3 Measure Classification

Now, with a clear concept of dimension, measure, and possible OLAP operations, we are ready to discuss implementation issues, i.e., how to compute the aggregated graph in a multi-dimensional and multi-level way.

Recall that in traditional OLAP, measures can be classified into *distributive*, *algebraic*, and *holistic*, depending on whether the measures of high-level cells can be easily computed from their low-level counterparts, without accessing base tuples residing at the finest level. For instance, in the classic *sale*(*time*, *location*) example, the total sale of [*2008*, *California*] can be calculated by adding up the total sales of [*January 2008*, *California*], [*February 2008*, *California*], …, [*December 2008*, *California*], without looking at base data points such as [*04/12/2008*, *Los Angeles*], which means that SUM is a distributive measure. Compared to this, AVG has been

used to illustrate algebraic measures, which is actually a *semi-distributive* category in that AVG can be derived from two distributive measures: SUM and COUNT, i.e., algebraic measures are functions of distributive measures.

(Semi-)distributiveness is a nice property for top-down cube computation, where the cuboid lattice can be gradually filled up by making level-by-level aggregations. Measures without this property is put into the holistic category, which is intuitively much harder to calculate. Now, concerning InfoNetOLAP, based on similar criteria with regard to the aggregated graphs, we can also classify them into three categories.

Definition 6 (Distributive, Algebraic, and Holistic) Consider a high-level cell c_h and the corresponding low-level cells it covers: $c_l^1, c_l^2, \ldots$. An aggregated graph M_d is *distributive* if $M_d(c_h)$ can be directly computed as a function of $M_d(c_l^1), M_d(c_l^2), \ldots$, i.e.,

$$M_d(c_h) = F_d\big[M_d(c_l^1), M_d(c_l^2), \ldots\big].$$

For a non-distributive aggregated graph M_a, if it can be derived from some other distributive aggregated graphs $M_d^1, M_d^2, \ldots$, i.e., for $\forall c_h$,

$$M_a(c_h) = F_a\big[M_d^1(c_h), M_d^2(c_h), \ldots\big],$$

then we say that it is *algebraic*. Aggregated graphs that are neither distributive nor algebraic belong to the *holistic* category.

If we further distinguish between I-OLAP and T-OLAP, there are actually six classes: I-distributive, I-algebraic, I-holistic and T-distributive, T-algebraic, T-holistic. Because of their wide differences with regard to semantics as well as implementation, let us first focus on InfoNet I-OLAP.

I-Distributive. The I-aggregated graph describing collaboration frequency in Example 1 is an I-distributive measure, because the frequency value in a high-level I-aggregated graph can be calculated by simply adding those in corresponding low-level I-aggregated graphs.

I-Algebraic. The graph displaying maximum flow configuration in Example 2 is an I-algebraic measure based on the following reasoning. Suppose *transportation type* and *company* comprise the two dimensions of a cube, and we generalize from low-level cells *(all-types, company 1)*, *(all-types, company 2)*, ... to *(all-types, all-companies)*, i.e., compute the maximum flow based on all types of transportation operated by all companies. Intuitively, this overall maximum flow would not be a simple sum (or other indirect manipulations) of each company's individual maximum flows. For instance, company 1 may have excessive transporting capability between two cities, whereas the same link happens to be a bottleneck for company 2: Considering both companies together for determination of the maximum flow can enable capacity sharing and thus create a double-win situation. In this sense, maximum flow is not distributive by definition. However, as an obvious fact, maximum flow f is decided by the network c that shows link capacities on its edges, and this capacity graph is distributive because it can be directly added upon generalization:

When link sharing is enabled, two separated links from A to B, which are operated by different companies and have capacities c^1_{AB} and c^2_{AB}, respectively, are no different from a single link with capacity $c^1_{AB} + c^2_{AB}$, considering their contributions to the flow value. Finally, being a function of distributive aggregated graphs, maximum flow is algebraic.

I-holistic. The I-holistic case involves a more complex aggregated graph, where base-level details are required to compute it. In the coauthor network of Example 1, the median of researchers' collaboration frequency for all DB conferences from 1990 to 1999 is holistic, similar to what we have seen in a traditional data cube.

Based on the same classification criteria, considering InfoNet T-OLAP, *T-distributive*, *T-algebraic*, and *T-holistic* aggregated graphs also exist. As an example, we will prove below that the measure graph showing degree centralities of vertices in a social network G is T-algebraic, where the *degree centrality* $C_D(v)$ of a vertex v is equivalent to the number of edges that connect v to other vertices in G.

Suppose we are performing topological roll-up operations on the coauthor network as we did in Fig. 16.2, and we are going from the individual level to the department level and then to the institution level. Here, we count each coauthored publication as an edge; so, if the collaboration frequency of two persons is 4, it means that there are four edges. Now, consider a set of vertices $V = \{v_1, v_2, \ldots, v_s\}$ that belong to the same node group and thus will be merged together into a single vertex v' by T-OLAP operations; it is easy to verify that

$$C_D(v') = \sum_{1 \le i \le s} C_D(v_i) - 2|E_V|,$$

where E_V is the number of edges that have both their ends in V. Figure 16.3 is an illustration of the computation process. In G'', it can be seen that $C_D(v')$ is a total of 4, 2, 5, and 3, which is 14. We can get this number directly from the original network G by the given formula $\sum_{1 \le i \le s} C_D(v_i) - 2|E_V| = (3+8+3+7+10+11+7+5+6) - 2(2+3+3+1+2+4+1+2+4+1) = 14$, while we can also leverage results obtained from the intermediate network G': $(8+12+14) - 2(3+1+6) = 14$. Since the computational cost is $O(s + |E_V|)$, where $s = |V|$ is the number of vertices in V, this example demonstrates that the computation cost can be greatly reduced by following the top-down materialization order and taking advantage of results that are already computed along the way, as G' has much less vertices than G.

16.4 Optimizations

Being (semi-)distributive or holistic tells us whether the aggregated graph computation needs to start from completely un-aggregated data or some intermediate results can be leveraged. However, even if the aggregated graph is distributive or algebraic, and thus we can calculate high-level measures based on some intermediate-level

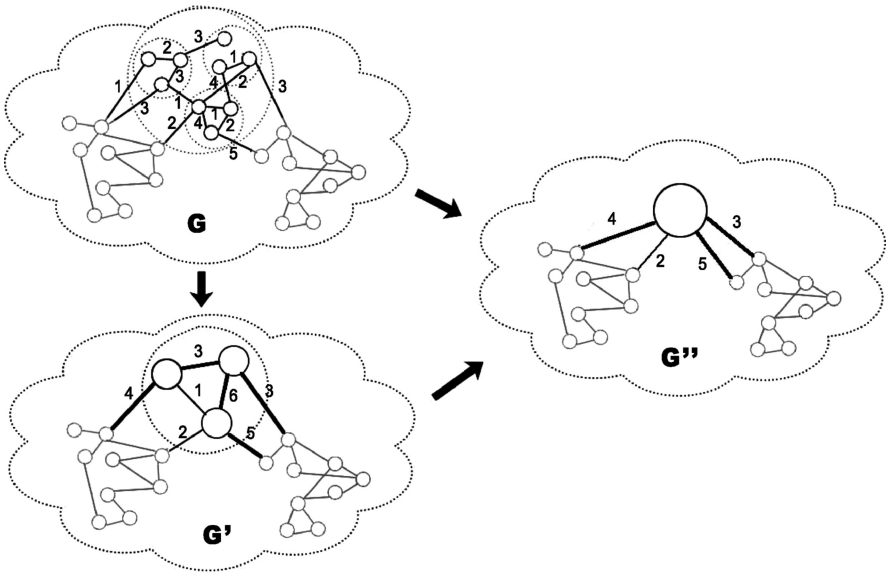

Fig. 16.3 Computing degree centrality

ones, it is far from enough, because the complexity to compute the two functions F_d and F_a in Definition 6 is another question. Think about the maximum flow example we just mentioned; F_a takes the distributive capacity graph as input to compute the flow configuration, which is by no means an easy transformation.

Based on our analysis, there are mainly two reasons for such potential difficulties. First, due to the interconnecting nature of graphs, the computation of many information network properties is "global" as it requires us to take the whole network into consideration. In order to make this concept of globalness clear, let us first look at a "local" situation: For I-OLAP, aggregated graphs are built for the same set of objects as the underlying network snapshots; now, in the aggregated graph of Example 1, "R. Agrawal" and "R. Srikant"'s collaboration frequency for cell *(db-conf, 2004)* is locally determined by "R. Agrawal" and "R. Srikant"'s collaboration frequency for each of the DB conferences held in 2004; it does not need any information from the other authors to fulfill the computation. This is an ideal scenario, because the calculations can be localized and thus greatly simplified. Unfortunately, not all measures provide such local properties. For instance, in order to calculate a maximum flow from S to T for the cell *(air, all-companies)* in Example 2, only knowing the transporting capability of each company's direct flights between S and T is not enough, because we can always take an indirect route via some other cities to reach the destination.

Second, the purpose for us to compute high-level aggregated graphs based on low-level ones is to reuse the intermediate calculations that are already performed. Taking I-OLAP for example, when computing the aggregated graph of a low-level cell c_l^i $(i = 1, 2, \ldots)$, we only had a partial view about the c_l^i-portion of network

snapshots; now, as multiple pieces of information are overlaid into the high-level cell c_h, some full-scale consolidation needs to be performed, which is very much like the merge sort procedure, where partial ranked lists are reused but somehow adjusted to form a full ranked list. Still, because of the structural complexity, it is not an easy task to develop such reuse schemes for information networks, and even if it is possible, reasonably complicated operations might be involved. Similar problems exist for InfoNet T-OLAP as well.

Admitting the difficulties above, let us now investigate the possibility to alleviate them. As we have seen, for some aggregated graphs, the first aspect can be helped by their *localization* properties with regard to network topology. Concerning the second aspect, the key is how to effectively reuse partial results computed for intermediate cells so that the workload to obtain a full-scale measure is *attenuated* as much as possible. In the following, we are going to examine these two directions in sequel.

16.4.1 Localization

Definition 7 (Localization in InfoNet I-OLAP) For an I-aggregated graph M_l^I that summarizes a group of network snapshots $\mathcal{G}' = \{\mathbb{G}_{i_1}, \mathbb{G}_{i_2}, \ldots, \mathbb{G}_{i_{N'}}\}$, if (1) we only need to check a *neighborhood* of v in $G_{i_1}, G_{i_2}, \ldots, G_{i_{N'}}$ to calculate v's node attributes in M_l^I, and (2) we only need to check a *neighborhood* of u, v in $G_{i_1}, G_{i_2}, \ldots, G_{i_{N'}}$ to calculate (u, v)'s edge attributes in M_l^I, then the computation of M_l^I is said to be *I-OLAP localizable*.

Example 4 (Common Friends) With regard to the coauthor network depicted in Example 1, we can also compute the following aggregated graph: Given two authors a_1 and a_2, the edge between them records the number of their common "friends", whereas in order to build such "friendship", the total collaboration frequency between two researchers must surpass a δ_c threshold for the specified conferences and time. ■

The above example provides another instance that leverages localization to promote efficient processing. Consider a cell, e.g., *(db-conf, 2004)*, the determination of the aggregated graph's a_1-a_2 edge can be restricted to a 1-neighborhood of these two authors in the un-aggregated snapshots of 2004's DB conferences, i.e., we only need to check edges that are directly adjacent to either a_1 or a_2, and in this way a third person a_3 can be found, if he/she publishes with both a_1 and a_2, while the total collaboration frequency summed from the weights of these adjacent edges is at least δ_c. Also, note that the above aggregated graph definition is based on the number of length-2 paths like a_1-a_3-a_2 where each edge of it represents a "friendship" relation; now, if we further allow the path length to be at most k, computations can still be localized in a $\lfloor \frac{k}{2} \rfloor$-neighborhood of both authors, i.e., any relevant author on such paths of "friendship" must be reachable from either a_1 or a_2 within $\lfloor \frac{k}{2} \rfloor$ steps. This can be seen as a situation that sits in the middle of Example 1's "absolute locality" (0-neighborhood) and maximum flow's "absolute globality" (∞-neighborhood).

There is an interesting note we want to put for the *absolutely local distributive* aggregated graph of Example 1. Actually, such a 0-neighborhood localization property degenerates the scenario to a very special case, where it is no longer necessary to assume the underlying data as an information network: For each pair of coauthors, we can construct a traditional cube showing their collaboration frequency "OLAPed" with regard to *venue* and *time*, whose computation does not depend on anything else in the coauthor network. In this sense, we can treat the coauthor network as a union of pairwise collaboration activities, whereas Example 1 can indeed be thought as a traditional OLAP scenario disguised under its information network appearances, because the cube we constructed here is nothing different from a collection of pairwise traditional cubes. As a desirable side effect, this enables us to leverage specialized technologies that are developed for traditional OLAP, which in general could be more efficient. Nevertheless, the case is special, anyway: Absolute localization would not hold for most information network measures, which is also the reason why traditional OLAP proves to be extremely restricted when handling network data.

The corresponding definition of localization in InfoNet T-OLAP is given as follows, and it can be easily proved that the degree centrality we have discussed above is 1-neighborhood T-OLAP localizable.

Definition 8 (Localization in InfoNet T-OLAP) For a T-aggregated graph M_l^T where a vertex v' of it represents a group of nodes $V = \{v_1, v_2, \ldots, v_s\}$ of the original network G_i, if (1) we only need to check a *neighborhood* of $v_1, v_2, \ldots$ in G_i to calculate v''s node attributes in M_l^T, and (2) we only need to check a *neighborhood* of $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ in G_i to calculate (u', v')'s edge attributes in M_l^T, then the computation of M_l^T is said to be *T-OLAP localizable*.

16.4.2 Attenuation

Below, we are going to explain the idea of attenuation through examples, and the case we pick is maximum flow. In a word, *the more partial results from intermediate calculations are utilized, the more we can decrease the cost of obtaining a full-scale aggregated graph*.

To begin with, let us first review some basic concepts, cf. [8]. Given a directed graph $G = (V, E)$, $c : \binom{V}{2} \to R^{\geq 0}$ indicates a *capacity* for all pairs of vertices and E is precisely the set of vertex pairs for which $c > 0$. For a source node s and a destination node t, a *flow* in G is a function $f : \binom{V}{2} \to R$ assigning values to graph edges such that (i) $f(u, v) = -f(v, u)$: skew symmetry, (ii) $f(u, v) \leq c(u, v)$: capacity constraint, and (iii) for each $v \neq s/t$, $\sum_{u \in V} f(u, v) = 0$: flow conservation. Since most maximum flow algorithms work incrementally, there is an important lemma as follows.

Lemma 1 *Let f be a flow in G and let G_f be its residual graph, where* residual *means that the capacity function of G_f is $c_f = c - f$; now, f' is a maximum flow in G_f if and only if $f + f'$ is a maximum flow in G.*

Note that the $+/-$ notation here means edge-by-edge addition/subtraction; and in summary, this lemma's core idea is to look for a flow f' in G_f and use f' to *augment* the current flow f in G.

For the InfoNetOLAP context we consider, in order to compute the algebraic aggregated graph displaying maximum flow, the function F_a takes a distributive capacity graph c as its input; now, since capacity can be written as the sum of a flow and a residual graph: $c = f + c_f$, does this decomposition provide us some hints to pull out the useful part f, instead of blindly taking c and starting from scratch?

Suppose that the capacity graph of cell *(all-types, company 1)* is c^1, where f_1 is the maximum flow and $c^1_{f_1} = c^1 - f_1$ denotes the corresponding residual graph. Likewise, we have c^2, f_2, and $c^2_{f_2}$ for cell *(all-types, company 2)*. Without loss of generality, assume there are only these two companies whose transportation networks are overlaid into *(all-types, all-companies)*, which has a capacity of $c = c^1 + c^2$.

Claim $f_1 + f_2 + f'$ is a maximum flow for c if and only if f' is a maximum flow for $c^1_{f_1} + c^2_{f_2}$.

Proof Since f_1 and f_2 are restricted to the transportation networks of company 1 and company 2, respectively, the overall capacity $c = c^1 + c^2$ must accommodate $f_1 + f_2$, even if link sharing is not enabled. As a result of subtracting $f_1 + f_2$, the residual graph becomes

$$\begin{aligned} c_{f_1+f_2} &= (c^1 + c^2) - (f^1 + f^2) \\ &= (c^1 - f^1) + (c^2 - f^2) = c^1_{f_1} + c^2_{f_2}. \end{aligned}$$

A direct application of Lemma 1 finishes our proof. ■

As it is generally hard to localize maximum flow computations with regard to network topology, the above property is important because it takes another route, which reuses partial results f_1, f_2 and attenuates the overall workload from c^1+c^2 to $c^1_{f_1}+c^2_{f_2}$. By doing this, we are much closer to the overall maximum flow f_1+f_2+f' because a big portion of it, $f_1 + f_2$, has already been decided even before we start an augmenting algorithm.

However, we should admit that attenuation schemes usually take widely different forms, which might need to be developed with regard to specific aggregate measure graphs; furthermore, as we shall see next, there do exist cases where such properties are hard, if not impossible, to think of.

Example 5 (Centrality) *Centrality* is an important concept in social network analysis, which reflects how "central" a particular node's position is in a given network. One definition called *betweenness centrality* C_B uses shortest path to model this: Let n_{jk} denote the number of shortest paths (as there could be equally short ones) between two nodes j and k; for any node i, $\frac{n_{jk}(i)}{n_{jk}}$ is the fraction of shortest paths between j, k that go through i, with $C_B(i)$ summing it up over all possible pairs: $C_B(i) = \sum_{j,k \neq i} \frac{n_{jk}(i)}{n_{jk}}$. Intuitively, for a "star"-shaped network, all shortest

paths must pass the network center, which makes C_B achieve its maximum value $(|V|-1)(|V|-2)/2$.

Only considering shortest paths is inevitably restrictive in many situations; and thus, *information centrality* C_I goes one step further by taking all paths into account. It models any path from j to k as a signal transmission, which has a channel noise proportional to its path length. For more details, we refer the readers to [25], which has derived the following formula based on information theoretic analysis: Let A be a matrix, whose a_{ij} entry designates the interaction strength between node i and node j; define $B = D - A + J$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{|V|} a_{ij}$ and J is a matrix having all unit elements; now, perform an inverse operation to get the *centrality matrix* $C = B^{-1}$; write its diagonal sum as $T = \sum_{j=1}^{|V|} c_{jj}$ and its row sum as $R_i = \sum_{j=1}^{|V|} c_{ij}$; the information centrality of node i is then equivalent to

$$C_I(i) = \frac{1}{c_{ii} + (T - 2R_i)/|V|}.$$

Now, with regard to the coauthor network described in Example 1, if we define the interaction strength between two authors as their total collaboration frequency for a set of network snapshots, then an aggregated graph M_{cen} can be defined, whose node i is associated with a node attribute $C_I(i)$ equivalent to its information centrality. ■

Claim The computation of M_{cen} is hard to be attenuated in a level-by-level aggregation scheme.

Proof As we can see, the core component of information centrality computation is a matrix inverse. Now, given two portions of network snapshots that are overlaid, the overall centrality matrix is

$$\big[(D_1 + D_2) - (A_1 + A_2) + J\big]^{-1} = (B_1 + B_2 - J)^{-1}.$$

From calculations performed on lower levels, we know the centrality matrices $C_1 = B_1^{-1}$ and $C_2 = B_2^{-1}$; however, it seems that they do not help much to decrease the computation cost of inverting $B_1 + B_2 - J$. ■

When things like this happen, an alternative is to abandon the exactness requirement and use intermediate results that are readily available to bound the answer within some range instead; as we shall elaborate in the following section, this will become very useful if the cube construction is subject to a set of constraints.

16.5 Constraints and Partial Materialization

Above, we have focused on the computation of a full cube, i.e., each cell in each cuboid is calculated and stored. In many cases, this is too costly in terms of both

space and time, which might even be unnecessary if the users are not interested in obtaining all the information. Usually, users may stick with an *interestingness* function I, indicating that only those cells above a particular threshold δ make sense to them. Considering this, all cells c with $I(c) \geq \delta$ comprise an *iceberg* cube, which represents a *partial materialization* of the cube's interesting part. Taking Example 2 for instance, it is possible that people may only care about those subnetworks that can transmit at least $\delta_{|f|}$ units of cargo, while the other cells are discarded from consideration, due to their limited usefulness for the overall transportation business.

Optimizations exist as to how such an iceberg cube can be calculated, i.e., how to efficiently process constraints like $I(c) \geq \delta$ during materialization, without generating a full cube at first. Below, we will first classify different constraints into categories (Section 16.5.1) and then combine with some examples to see how each category should be dealt with (Section 16.5.2).

16.5.1 Constraint Classification

Two most important categories of constraints are *anti-monotone* and *monotone*. They relate cells on different levels of the cube together and are defined as follows.

Definition 9 A constraint C is *anti-monotone*, if for any high-level cell c_h and a low-level cell c_l covered by c_h, the following must hold: c_h violates $C \Rightarrow c_l$ violates C.

Definition 10 A constraint C is *monotone*, if for any high-level cell c_h and a low-level cell c_l covered by c_h, the following must hold: c_h satisfies $C \Rightarrow c_l$ satisfies C.

Note that, in Definition 9, "c_h violates $C \Rightarrow c_l$ violates C" is equal to "c_l satisfies $C \Rightarrow c_h$ satisfies C" and, in Definition 10, "c_h satisfies $C \Rightarrow c_l$ satisfies C" is equal to "c_l violates $C \Rightarrow c_h$ violates C"; so, depending on whether c_h or c_l is computed (and thus verified against C) first, there are different ways to make use of these constraints, which we will demonstrate below.

16.5.2 Constraint Pushing

The anti-monotone and monotone constraints can be "pushed" deep into the computation process using the Apriori principle [30]. In general, there are two approaches to compute a data cube over information networks, *bottom-up* and *top-down*. In bottom-up computation, which can be contrasted with BUC [2] in traditional OLAP, high-level cells are calculated first, before drilling down to low-level cells they cover. In top-down computation, which can be contrasted with Multi-Way [31] in traditional OLAP, we calculate low-level cells first, and then aggregate to high-level cells. Finally, which approach to adopt will depend on various parameters, including the size of the network, data sparsity, the measures to be computed, and the available constraints.

Now consider the bottom-up approach; on the one hand, if a high-level cell c_h does not satisfy an anti-monotone constraint, then we know that no low-level cell c_l covered by c_h would satisfy it, and thus the calculation can be immediately terminated, pruning c_l and its descendants from the cuboid lattice; on the other hand, if a high-level cell c_h already satisfies a monotone constraint, then we no longer need to perform checkings for any low-level cells covered by c_h because they would always satisfy it. As for top-down computations, the roles of anti-monotonicity and monotonicity are reversed accordingly.

It is easy to see that anti-monotone and monotone properties depend on specific analysis of measures and interestingness functions. Here, since we are working with network data, some graph theoretic studies need to be made. Let us examine a few examples.

Claim Suppose maximum flow is the I-OLAP measure to be calculated, regarding its flow's value $|f| = \sum_{v \in V} f(s, v) = \sum_{v \in V} f(v, t)$, i.e., the highest amount of transportation a network can carry from s to t, constraint $|f| \geq \delta_{|f|}$ is anti-monotone.

Proof Intuitively, the transporting capability of one company must be smaller than that of all companies together, since there are now more available links for the flow to pass. In fact, as we showed in Section 16.4, the flow value of c_h is no smaller than the flow sum of *all* c_l's that are covered by c_h, which is a condition stronger than the normal anti-monotonicity defined between a high-level cell and a *single* low-level cell it covers. ■

Claim The diameter of an information network G is designated as the maximum shortest path length for all pairs of nodes in G. Now, denote diameter as d and let it be the T-OLAP measure we want to calculate, constraint $d \geq \delta_d$ is monotone.

Proof As we perform roll-up in InfoNet T-OLAP and shrink the network, some nodes are merged into one during this process, which can only shorten relevant paths. ■

Because of space limit, we are not going to list more examples here. Certainly, having the conditions on interestingness classified into anti-monotone and monotone will greatly help us in constructing an iceberg graph cube. Here, combined with top-down/bottom-up calculations, Apriori is a simple and effective principle to align search space pruning with particular computation orders. But unfortunately, we cannot use it to push every constraint into the mining process, as there do exist situations that are neither anti-monotone nor monotone. To this extent, we shall further investigate how the other kinds of more complex constraints can be nicely handled in the future.

16.6 Discovery-Driven InfoNetOLAP

Apart from the type of data being dealt with, another important distinction of InfoNetOLAP from the traditional one is the need for flexibility in manipulating

information networks. Thus, in addition to the uniform drilling of traditional OLAP, where all nodes at a given level of abstraction are simultaneously rolled up or drilled down, e.g., from month to quarter, and then to year, InfoNetOLAP may require *selective drilling* for a given node or neighborhood. For instance, in the coauthor network example, a user could be interested in the collaborative relationship between Yahoo! Labs and related individual researchers. Such an analysis may show strong collaborations between AnHai Doan at Wisconsin and people at Yahoo! Labs. So, if all Wisconsin researchers are merged into a single institution in the same way as Yahoo! Labs, it would be hard to discover such a relationship since, collectively, there would be even stronger collaborations between Wisconsin and Berkeley (instead of Yahoo! Labs), which may overshadow AnHai's link to Yahoo! Labs.

Selective drilling, though promising, may generate an exponential number of combinations, which are too costly to compute and explore. To ensure that one can pinpoint to the "real gold" during exploration, *discovery-driven InfoNetOLAP* should be adopted, i.e., rather than searching the complete cube space, one has to be sure that the planned drilling should help the discovery of interesting knowledge.

16.6.1 Discovery-Driven: Guiding Principles

In the following, we first outline a few guiding principles for discovery-driven InfoNetOLAP.

Discovery Driven by *Rule-based Heuristics*. When performing InfoNetOLAP, we usually have some general intuition about where and when the selective drilling should be focused and at what levels/nodes one is most likely to discover something interesting. For instance, when studying research collaborations, it may not be wise to quickly merge prominent researchers into groups, before exploring some interesting patterns for them at first. As an example, if Raghu Ramakrishnan is merged into an even "bigger" entity like Yahoo! Labs, it may blur the individual relationships that can be discovered; rather, we could keep his own identity in the network, on which clearer patterns with finer granularity are observed. Alternatively, for ordinary graduate students, it seems to be more interesting to group them together or have them absorbed by nearby "hub" nodes, because for such individuals, it is not likely that something significant can stand out from the rest. In this sense, a rule like *delay the merge of 'big' or 'distinct' nodes* could be quite simple to work out and follow, as the system only needs to consult attributes of an entity itself (e.g., how many papers a researcher has published) or explore some very local information (e.g., how many persons he/she has collaborated with) for decision making.

Discovery Driven by *Statistical Analysis*. In many cases, it is beneficial to conduct some global (rather than local) preliminary analysis before choosing any drilling operation. Consider a top-down exploratory scenario where the granularity is currently set at the institution level, e.g., both Yahoo! Labs and University of Wisconsin are treated as a whole; now, if an automatic background computation shows that

the collaboration activities between Yahoo! Labs and AnHai Doan are significantly higher than normal, then it is rewarding to drill-down, because the status of AnHai is like an outlier in Wisconsin; otherwise, such drilling may not disclose anything truly interesting. This is similar to discovery-driven OLAP proposed by [23], but in the environment of information networks. As a simple implementation of this idea, we may take the collaboration frequencies between Yahoo! Labs and every person from Wisconsin, and calculate the variance: If the variance is high, one may set up an indicator which may suggest people to click on "U. of Wisc." and expand it to get a refined view, which is just the situation described in Fig. 16.4. Compared to the first guiding principle, there are no longer simple rules that stipulate whether a drill-down or roll-up should be performed; everything depends on a global statistical analysis about the entities in the network, which aims to find out those interesting (or outlying) phenomena that are worthwhile to explore.

Discovery Driven by *Pattern Mining*. Another way for discovery-driven OLAP is fueled by the potential to discover interesting patterns and knowledge on information networks using data mining techniques: If splitting or merging of certain sets of nodes may lead to the discovery of interesting clusters, frequent patterns, classification schemes, and evolution regularities/outliers, then the drilling should be performed, with the results/patterns demonstrated to users. Otherwise, the drilling will not be performed. Notice that, although such pattern discovery can be executed on the fly at the time of user interaction, the discovery process could be too time-consuming with regard to the user's mouse clicking. Therefore, we suggest the pre-computation of some promising drilling paths as intermediate results to speed up interactive knowledge discovery. It is an interesting research issue to determine the things to be computed in order to facilitate such on-line analysis.

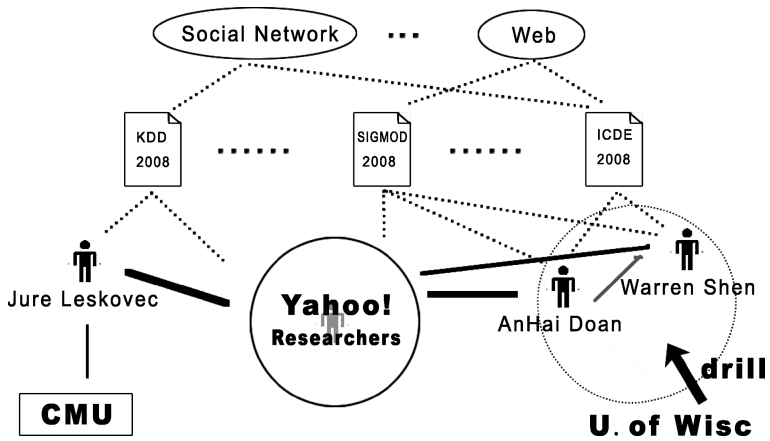

Fig. 16.4 Discovery-driven InfoNetOLAP

16.6.2 Network Discovery for Effective OLAP

As we discussed above, for effective discovery-driven OLAP, it is important to perform some essential data mining on the underlying information networks to reveal interesting network patterns that may help us discover the hidden knowledge. Here, we discuss some interesting network discovery procedures by taking the collaboration scenario of researchers as an example.

For studying coauthor networks, it is important to distinguish different roles the authors may play in the network. For example, based on the coauthor relationships over time, it is possible to dig out advisor–advisee relationships. Usually, advisee is a mediocre node in the network without many publications, who then starts to coauthor substantially with his prominent advisor; after graduation, he/she joins industry or moves on to another institution, and the publishing behaviors are changed again. Advisors can also be identified, based on his/her long-term publication history as well as a center role of working with many junior coauthors. It is interesting to organize researchers under such a phylogeny tree and examine the interactions between different clusters of academic "families".

Besides some not-so-sophisticated knowledge discovery procedures, a mining process may involve induction on the entire information networks as well. For example, in order to partition an interconnected, heterogeneous information network into a set of clusters and rank the nodes in each cluster, one could develop a RankClus framework [26], which integrates clustering and ranking together to effectively cluster information networks into multiple groups and rank nodes in each group based on certain nice properties (such as authority). By examining authors, research papers, and conferences, one can group conferences in the same fields together to form conference clusters, group authors based on their publication records into author clusters, and in the meantime rank authors and conferences based on their corresponding authorities. Such clustering results enhanced by ranking information would be an ideal feed into InfoNetOLAP. Interestingly, such clustering-ranking can be performed based on the links only, without checking the citation information and the keywords or text information contained in the conferences and/or publication titles. The details of such techniques are beyond the discussions of this chapter, but it sheds light on automated processes to effectively identify concept hierarchies and important nodes for discovery-driven InfoNetOLAP.

16.7 Experiments

In this section, we present empirical studies evaluating the effectiveness and efficiency of the proposed InfoNetOLAP framework. It includes two kinds of data sets, one real data set and two synthetic data sets. All experiments are done on a Microsoft Windows XP machine with a 3 GHz Pentium IV CPU and 1GB main memory. Programs are compiled by Visual C++.

16.7.1 Real Data Set

The first data set we use is the DBLP Bibliography (http://www.informatik.uni-trier.de/~ley/db/) downloaded in April 2008. Upon parsing the author field of papers, a coauthor network with multiple snapshots can be constructed, where an edge of weight w is added between two persons if they publish w papers together. We pick a few representative conferences for the following three research areas:

- Database (DB): PODS/SIGMOD/VLDB/ICDE/EDBT
- Data Mining (DM): ICDM/SDM/KDD/PKDD
- Information Retrieval (IR): SIGIR/WWW/CIKM

and also distribute the publications into 5-year bins: (2002, 2007], (1997, 2002], (1992, 1997],.... In this way, we obtain two informational dimensions: *venue* and *time*, on which I-OLAP operations can be performed.

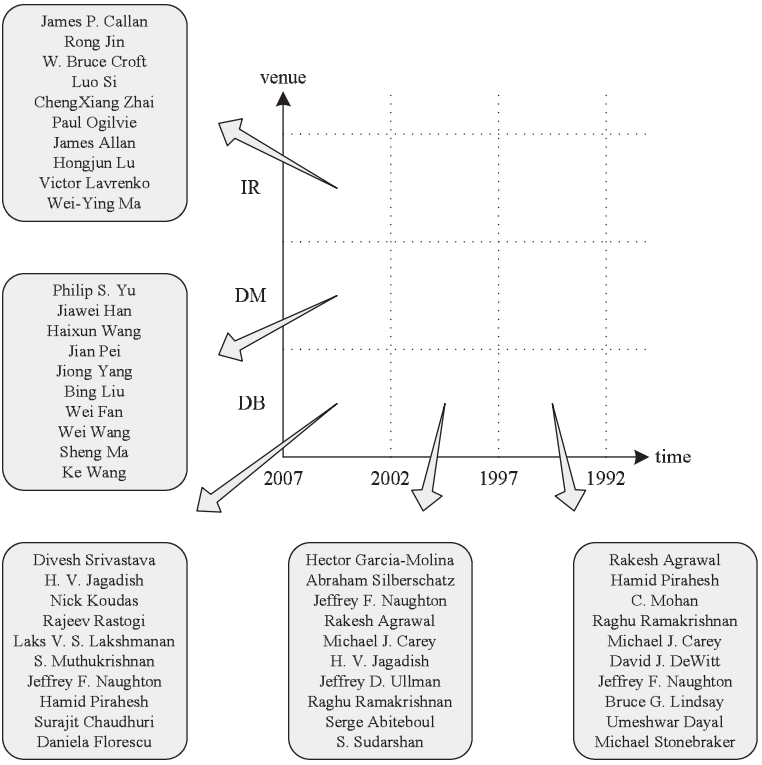

Fig. 16.5 A multi-dimensional view of top-10 "Central" authors

Figure 16.5 shows a classic OLAP scenario. Based on the definition in Section 16.4, we compute the information centrality of each node in the coauthor network and rank them from high to low. In general, people who not only publish a

lot but also publish frequently with a big group of collaborators will be ranked high. Along the *venue* dimension, we can see how the "central" authors change across different research areas, while along the *time* dimension, we can see how the "central" authors evolve over time. In fact, what Fig. 16.5 gives is a multi-dimensional view of the cube's base cuboid; without any difficulty, we can also aggregate DB, DM, and IR into a broad Database field, or generalize the *time* dimension to *all-years*, and then compute respective I-OLAP cells. Given each author's affiliation information, we may conduct T-OLAP and obtain most "central" research groups and institutions as well. The results are omitted here.

16.7.2 Synthetic Data Sets

We use synthetic data sets to demonstrate the effectiveness of the optimizations that are proposed to efficiently perform OLAP operations over information networks. The first test we pick is the computation of maximum flow as an InfoNetI-OLAP measure, which has been used as an exemplifying application in our discussions.

Generator Mechanism. Since it is generally hard to get real flow data, we develop a synthetic generator by ourselves. The data is generated as follows: The network has a source node s and a destination node t, and in between them, there are L intermediate layers, with each layer containing H nodes. There is a link with infinite capacity from s to every node in layer 1, and likewise from every node in layer L to t. Other links are added from layer i to layer $i + 1$ on a random basis: For the total number of $H \cdot H$ choices between two layers, we pick αH^2 pair of nodes and add a link with capacity 1 between them.

For the cube we construct, there are d dimensions; each dimension has $card$ different values (i.e., cardinality), which can be generalized to "all/*". For a base cell where all of its dimensions are set on the finest ungeneralized level, we generate a snapshot of capacity network $L5H1000\alpha 0.01$, i.e., there are five layers of intermediate nodes, and $0.01 \cdot (1000)^2 = 10,000$ links are randomly added between neighboring layers.

The algorithm we use to compute the maximum flow works in an incremental manner. It randomly picks an augmenting path from s to t until no such paths exist. To accommodate top-down computation, where high-level cells are computed after low-level cells so that intermediate results can be utilized, we integrate our attenuation scheme with the classic Multi-Way aggregation method for cube computation [31].

The results are depicted in Figs. 16.6 and 16.7, with Fig. 16.6 fixing the cardinality as 2 and varying d from 2, 3, ..., up to 6, and Fig. 16.7 fixing the number of dimensions as 2 and varying $card$ from 2, 4 ..., up to 8. It can be seen that the optimization achieved through attenuation is obvious, because in effect we do not need to compute a full-scale aggregated graph from scratch, and part of the burden has been transferred to previous rounds of calculations. Especially, when the dimensionality goes high in Fig. 16.6, so that more levels of cube cells are present, the superiority of attenuation-based methods becomes more significant, and one may reap orders of magnitude savings.

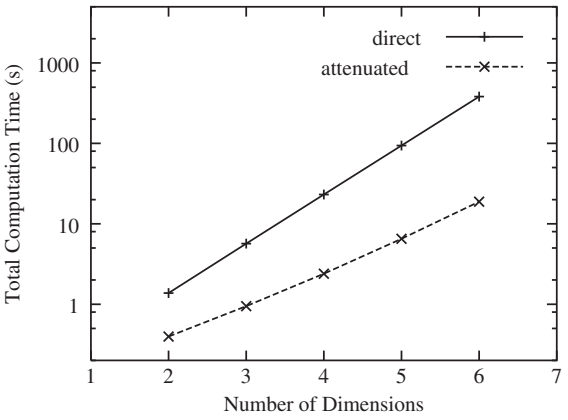

Fig. 16.6 The effect of optimization w.r.t. number of dimensions

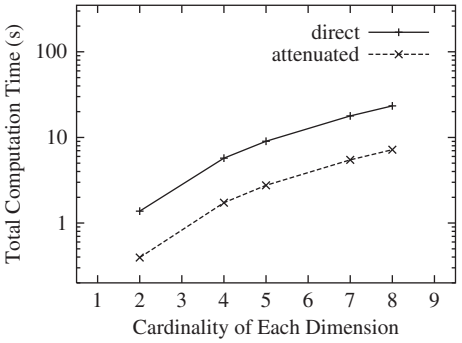

Fig. 16.7 The effect of optimization w.r.t. dimension cardinality

The second test we perform is on the calculation of degree centrality with regard to InfoNet T-OLAP.

Generator Mechanism. The synthetic data networks in this experiment are generated in the following manner: For a network G, n vertices are generated first, i.e., $|V(G)| = n$; as the next step, edges are randomly attached to vertices such that (1) the entire network is connected, (2) the vertices have an average degree of $\overline{d}$, and (3) the edges have an average weight of $\overline{w}$.

Given a network G, users can choose a subset of vertices $V \subseteq V(G)$ to merge into a single vertex v' and compute the T-aggregated graph for the generalized network G'. Such a roll-up operation is called a user T-OLAP request. For a network G, we recursively partition G into π non-overlapping connected components with equal number of vertices, e.g., suppose $|V(G)| = 1024$ and $\pi = 4$, we can first partition G into 4 connected subgraphs with 256 vertices, and then recursively partition these 4 subgraphs until there is only one vertex left. Finally, if we reverse the above sequence of partitioning, a series of T-OLAP requests will be naturally formed, and

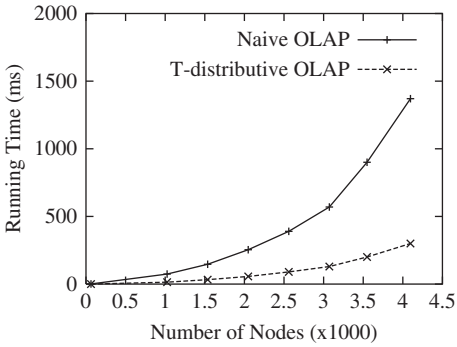

Fig. 16.8 Running time w.r.t. network size

our task next is to compute the corresponding aggregated graphs along the drilling path.

We refer to the baseline algorithm for comparison as Naive OLAP. For each T-OLAP request, this method would directly compute the T-aggregated graph from the original network G. As shown in Section 16.3, degree centrality is T-algebraic: Therefore, we can leverage the (semi-)distributiveness property and make use of the degree centralities that have been calculated for lower-level intermediate networks, which can offer significant efficiency boost. This computation strategy is denoted as T-distributive OLAP.

We set the average vertex degree as $\overline{d} = 5$, and fix the number of partitions per step as $\pi = 4$. Figure 16.8 shows the running time comparison for the two alternative approaches as the number of vertices in G increases. Here, the running time is the total computation cost summed over all T-OLAP requests, and it can be easily observed that with T-distributiveness the computation cost increases much slower than Naive OLAP. In Fig. 16.9, we fix $V(G)$ as 4096 and vary π, which depicts how the granularity of T-OLAP operations can affect both approaches: As π increases, the granularity difference between adjacent levels of networks becomes larger. Since

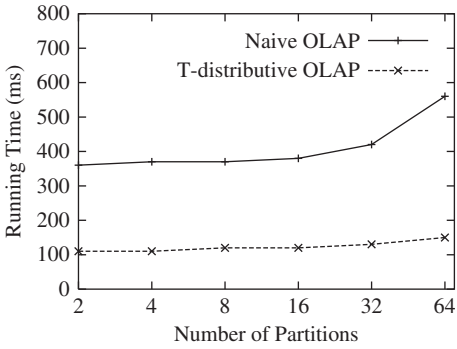

Fig. 16.9 Running time w.r.t. T-OLAP granularity

the computation of degree centrality is not too complex, both approaches have rather slow increase with regard to the running time, as we can see from the picture. However, notice that the T-distributive OLAP still features a flatter growth curve compared with that of the Naive OLAP approach.

16.8 Related Work

OLAP (On-Line Analytical Processing) is an important notion in data mining, which has drawn a lot of attention from the research communities. Representative studies include [7, 11], and a set of papers on materialized views and data warehouse implementations are collected in [12]. There have been a lot of works that deal with the efficient computation of a data cube, such as [2, 31], whereas the wealth of literature cannot be enumerated. However, all these researches target conventional spreadsheet data, i.e., OLAP analysis is performed on independent data tuples that mathematically form a set. In contrast, as far as we know, ours is the first that puts information networks in a rigid multi-dimensional and multi-level framework, where due to the nature of the underlying data, an OLAP measure in general takes the form of an aggregated graph.

The classification of OLAP measures into distributive, algebraic, and holistic was introduced in the traditional OLAP arena, where we can also find related works for iceberg cubing [9], partial materialization [17], and constraint pushing [21]. It is important to see how these basic aspects are dealt with in the InfoNetOLAP scenario; and as we have seen from the discussions, things become much more complicated due to the increased structural complexity.

In InfoNetOLAP, the aggregated graph can be thought as delivering a summarized view of the underlying networks based on some particular perspective and granularity, which helps users get informative insights into the data [6]. In this sense, concerning the generation of summaries for information networks, there have been quite a few researches that are associated with terminologies like compression, summarization, and simplification. For example, Boldi and Vigna [3, 22] study the problem of compressing large graphs, especially Web graphs; however, they only focus on how the Web link information can be efficiently stored and easily manipulated to facilitate computations such as PageRank and authority vectors, which do not provide any pointers into the network structures. Similarly, Beyer and Ramakrishnan [2] develop statistical summaries that analyze simple characteristics like degree distributions and hop-plots on information networks; Lu et al. [18] go one step further by looking into the evolutionary behavior of these statistics and proposes a generative model that helps explain the latent mechanism. These compressed views are useful but hard to be navigated with regard to the underlying networks; also, the multi-dimensional functionality that can conduct analysis from different angles is missing. Another group of papers [1, 15, 19, 29] are often referred as graph simplification, e.g., Archambault et al. [1] aim to condense a large network by preserving its skeleton in terms of topological features, and Kossinets et al. [15] try to extract the “backbone” of a social network, i.e., the subnetwork that consists of edges on

which information has the potential to flow the quickest. In this case, attributes on nodes and edges are not important, and the network is indeed an unlabeled one in its abstract form. Works on graph clustering (to partition similar nodes together), dense subgraph detection (for community discovery, link spam identification, etc.), graph visualization, and evolutionary pattern extraction/contrast include [20], [10], [13, 28], and [5] respectively. They all provide some kind of summaries, but the objective and result achieved are substantially different from those of this chapter.

With regard to summarizing attributed networks that incorporates OLAP-style functionalities, [25] is the closet to ours in spirit. It introduces an operation called SNAP (Summarization by grouping Nodes on Attributes and Pairwise relationships), which merges nodes with identical labels (actually, it might not be necessary to require exactly the same label for real applications, e.g., Lu et al. [18] introduce a way to find similar groups of entities in a network, and this can be taken as the basis to guide node merges), combines corresponding edges, and aggregates a summary graph that displays relationships for such "generalized" node groups. Users can choose different resolutions by a k-SNAP operation just like rolling up and drilling down in an OLAP environment. This can be seen as a special instance of InfoNet T-OLAP that is defined in this chapter.

Sarawagi et al. [23] introduce the discovery-driven concept for traditional OLAP, which aims at pointing out a set of direct handles or indirect paths that might lead to the interesting/outlying cells of a data cube. Their method is statistics oriented. As we proposed in this chapter, InfoNetOLAP can also adopt the discovery-driven concept, but in the context of information networks, which suggests new classes of discovery-driven methods using topological measures and network patterns. Different from traditional OLAP, where dimensions are often globally drilled down and rolled up, InfoNetOLAP takes selective drilling into consideration, which leverages graph mining (e.g., [14]), link analysis (e.g., [24]), etc., and might only involve some local portion of a big network.

16.9 Conclusions

We examine the possibility to apply *multi-dimensional* analysis on information networks, and develop an *InfoNetOLAP framework*, which is classified into two major subcases: *informational OLAP* and *topological OLAP*, based on the different OLAP semantics. Due to the nature of the underlying data, an OLAP measure now takes the form of an *aggregated graph*. We categorize aggregated graphs based on the difficulty to compute them in an OLAP context and suggest two properties: *localization* and *attenuation*, which may help speed up the processing. Both full materialization and constrained partial materialization are discussed. Toward more intelligent InfoNetOLAP, we further propose a *discovery-driven* multi-dimensional analysis model and discuss many challenging research issues associated with it. Experiments show insightful results on real data sets and demonstrate the efficiency of our proposed optimizations.

As for future works, there are a lot of directions we want to pursue on this topic, for example, extending the current framework to heterogeneous-typed information networks, hyper-graphs, etc., and our immediate target would be refining the discovery-driven InfoNetOLAP idea and testing it on several interesting application domains.

References

1. D. Archambault, T. Munzner, and D. Auber. Topolayout: Multilevel graph layout by topological features. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):305–317, 2007.
2. K. S. Beyer and R. Ramakrishnan. Bottom-up computation of sparse and iceberg cubes. In *SIGMOD Conference*, pages 359–370, 1999.
3. P. Boldi and S. Vigna. The WebGraph framework I: Compression techniques. In *WWW*, pages 595–602, 2004.
2. D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, 38(1), 2006.
5. J. Chan, J. Bailey, and C. Leckie. Discovering correlated spatio-temporal changes in evolving graphs. *Knowledge and Information Systems*, 16(1):53–96, 2008.
6. V. Chandola and V. Kumar. Summarization – compressing data into an informative representation. *Knowledge Information Systems*, 12(3):355–378, 2007.
7. S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26(1):65–74, 1997.
8. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2001.
9. M. Fang, N. Shivakumar, H. Garcia-Molina, Rajeev Motwani, and Jeffrey D. Ullman. Computing iceberg queries efficiently. In *VLDB*, pages 299–310, 1998.
10. D. Gibson, R. Kumar, and A. Tomkins. Discovering large dense subgraphs in massive graphs. In *VLDB*, pages 721–732, 2005.
11. J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Mining and Knowledge Discovery*, 1(1):29–53, 1997.
12. A. Gupta and I. S. Mumick. *Materialized Views: Techniques, Implementations, and Applications*. MIT Press, Cambridge, MA, 1999.
13. I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.
14. G. Jeh and J. Widom. Mining the space of graph properties. In *KDD*, pages 187–196, 2004.
15. G. Kossinets, J. M. Kleinberg, and D. J. Watts. The structure of information pathways in a social communication network. In *KDD*, pages 435–443, 2008.
18. J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.
17. X. Li, J. Han, and H. Gonzalez. High-dimensional olap: A minimal cubing approach. In *VLDB*, pages 528–539, 2004.
18. W. Lu, J. C. M. J., E. E. Milios, N. Japkowicz, and Y. Zhang. Node similarity in the citation graph. *Knowledge and Information Systems*, 11(1):105–129, 2007.
19. S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In *SIGMOD Conference*, pages 419–432, 2008.
20. A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

21. R. T. Ng, L. V. S. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rules. In *SIGMOD Conference*, pages 13–24, 1998.
22. S. Raghavan and H. Garcia-Molina. Representing web graphs. In *ICDE*, pages 405–416, 2003.
23. S. Sarawagi, R. Agrawal, and N. Megiddo. Discovery-driven exploration of olap data cubes. In *EDBT*, pages 168–182, 1998.
24. P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad. Collective classification in network data. *CS-TR-4905, University of Maryland, College Park*, 2008.
25. K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11(1):1–37, 1989.
26. Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: Integrating clustering with ranking for heterogeneous information network analysis. In *EDBT*, pages 565–576, 2009.
25. Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *SIGMOD Conference*, pages 567–580, 2008.
28. N. Wang, S. Parthasarathy, K.-L. Tan, and A. K. H. Tung. Csv: visualizing and mining cohesive subgraphs. In *SIGMOD Conference*, pages 445–458, 2008.
29. A. Y. Wu, M. Garland, and J. Han. Mining scale-free networks using geodesic clustering. In *KDD*, pages 719–724, 2004.
30. X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. F. M. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. S., D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.
31. Y. Zhao, P. Deshpande, and J. F. Naughton. An array-based algorithm for simultaneous multidimensional aggregates. In *SIGMOD Conference*, pages 159–170, 1997.