Authentication for Stock Data Prediction using Map Reduce in Big Data

Mr. S. Nagaraj[1] Kavin P[2] Manikandan T[3] Nalankilli R[4] Raghul V[5]
Assistant Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India 1Student, Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India 2,3,4,5

ABSTRACT- In the context of large data, this study introduces a unique method for user authentication combining Directed Acyclic Graphs (DAG) and Frequent Item set Utility Frequency Pattern Analysis. Securing accessto sensitive data is crucial in today's data-driven environment, and handling large-scale datasets may make standard authentication techniques less effective. The suggested system effectively handles enormous volumes of authentication data by utilizing the capabilities of DAG-based processing. Parallel processing is made possible by segmenting the authentication procedure into directed acyclic subgraphs, which greatly shortens the computation time. Additionally, the system can recognize recurrent authentication patterns with related utility values thanks to the addition of Frequent Item set Utility Frequency Pattern Analysis. This improves overall security and offers useful insights for anomaly identification.
Keywords: Big Data, Graph, Map reduce

I.INTRODUCTION

1.1 GRAPH

A graph is a mathematical depiction of a collection of items, referred to as vertices or nodes, and the edges that connect them. In many disciplines, including computer science, mathematics, the social sciences, and more, graphs are used extensively to model and examine interactions between distinct elements. Graphs can be classified as directed or undirected, depending on whether edges have a defined direction or none at all. They can also be unweighted, meaning that the edges have no associated values, or weighted, meaning that the edges contain numerical values. Individual entities are represented by the vertices of a graph,
The relationships, interactions, or connections between those entities are represented by the edges. In a social network graph, for instance, friends or connections between users would be represented as edges, and individuals themselves would be represented as vertices. The study of graphs is known as graph theory, and it encompasses a number of concepts and algorithms for effectively analyzing and processing graphs. Identifying cycles, computing connectivity components, determining shortest pathways, and resolving network flow and optimization issues are a few typical graph algorithms. Graphs are a vital tool in many different applications, such as recommendation systems, computer network routing, social network analysis, logistics and transportation planning, and many more. They offer a strong and versatile means ofrepresenting intricate interactions and structures.

1.2 BIG DATA

Fig 1 Process of Big data

A key component of contemporary data-driven analysis and decision-making is big data. Big data can be utilized by researchers and organizations to find trends, patterns, and correlations that improve knowledge and process optimization. Big data analytics is very important to industries like marketing, banking, e- commerce, healthcare, and transportation since it helps them make data-driven decisions and obtain a competitive edge. Specialized tools

and technologies, notably distributed computing frameworks like Apache Hadoop and Apache Spark, have been developed to handle large data. These frameworks offer effective handling of large-scale data collections by enabling distributed data processing and storage over a cluster of devices. Big data has many benefits, but managing and analyzing it can be difficult at times.

1.3 MAPREDUCE

The Map Reduce is fault-tolerant that is, it can withstand node outages and keep processing without losingdata it is built to manage massive amounts of data processing. The intricacies of fault tolerance, data distribution, and parallel processing are abstracted away, allowing developers to concentrate more easily on crafting the "map" and "reduce" functions that correspond to their particular data processing needs. Theopen-source Apache Hadoop system, which enables developers to handle enormous datasets across clusters of commodity hardware, is the most well-known example of a Map Reduce implementation. Hadoop offersa Map Reduce engine for data processing and a distributed file system (HDFS) for data storage. Despite itspopularity and strength, Map Reduce is not appropriate for all kinds of data processing jobs. Certain problems may not fit well into the map-reduce paradigm, and the overhead of several Map Reduce phasesmay cause performance issues for iterative methods. Consequently, in order to overcome some of these constraints and provide more adaptable and effective methods for working with big data, alternative data processing frameworks have been created, such as Apache Spark.

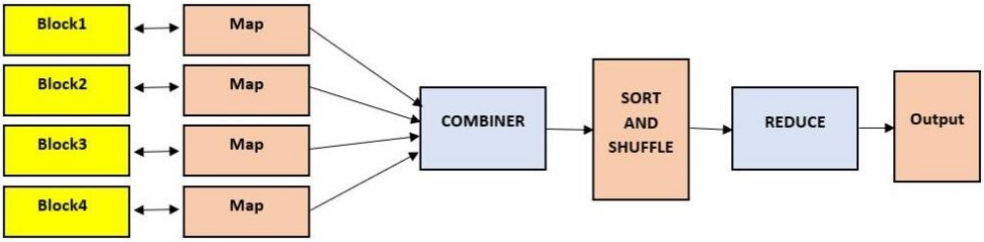

Fig 2. MAPREDUCE

2. LITERATURE REVIEW

2.1 THEORY OF GRAPHS AND ITS APPLICATIONS

George Mohler [1]et al. As this study suggests, a diagram comprising a collection of dots and lines connecting specific pairings of these points can effectively depict numerous real-world scenarios. Points might be used to represent persons, for instance, with lines connecting friends in pairs, or they could be used to represent communication centers, with lines signifying links between conversations. Note that the major focus of these diagrams is whether or not two specified points are connected by a line; the specific method of connection is irrelevant. This kind of circumstance can be mathematically abstracted to create the concept of a graph. The goal of this book is to provide an overview to graph theory. Our goal has been to provide what we believe to be the foundational knowledge together with an extensive range of applications, to other areas of mathematics as well as to practical issues. Simple new proofs of Brooks, Chviital, Tuttle, and Vizing's theorems are included. The applications are discussed in significant detail andhave been carefully chosen.

2.2 A LINEAR ALGEBRAIC METHODS TO SOME GRAPH THEORY PROBLEMS

Elizabeth A. Kalinina [2] and associates. As suggested in this paper, we examine a few well-known graph theory issues from the perspective of linear algebra. By examining the characteristics of vector spaces overGF (2), we can verify the theorem concerning graph circuits and cut-sets and create an innovative approachfor identifying line graphs and creating their original graphs. This work examines n-dimensional vector spaces over the field GF (2), which has two elements, 0 and 1. Addition and multiplication are performed using standard integer operations, while reduction is performed modulo 2. We demonstrate a new approachto identify a line graph and create its original graph, and we prove a theorem on graph circuits and cut-sets with the aid of unique features of these vector spaces. Given that several real-world applications of graph theory have relatively straightforward answers for line graphs, line graph

recognition is a significant problem. A linear algebraic approach to graph analysis yields numerous intriguing findings. The renowned Cheeger's inequality approximating the sparsest cut-set, for instance, is one of the most valuable facts in algorithmic applications. Theorems linking graph diameter and eigenvalues were also proven.

2.3 K-CONNECTIVITY TESTING OF DIRECTED AND UNDIRECTED GRAPHS USING FASTPARALLEL ALGORITHMS

Liang Weifa [3] et al. It appears that no NC algorithms, even for fixed k > 1, have ever been developed forevaluating a directed graph for k-edge or k-vertex connection. We provide such algorithms with a time complexity of O(k1ogn) employing nP(n,m) or (n+k2)P(n,m) processors, respectively, by the use of a basicflow approach. In a directed graph with O(n) vertices and O(m) edges, where P(n,m) is the number of processors needed to find some path in time O(1og n) between two specified vertices; the computation model is a CRCW PRAM. These algorithms, of course, also apply to undirected graphs, but for both typesof connectivity we can improve the factors P(n,m) to P(n,kn) by using sparse certificates. This algorithm for undirected graphs is faster than earlier ones by a factor of O(k) in terms of time. Furthermore, even in the case where k is not constant, edge connection is NC-reducible to vertex connectivity. This work employs the concurrent read and concurrent write parallel random access machine (CRCW PRAM) concept of parallel computing. Multiple processors can access the same memory location simultaneously in this approach for both read and write operations. An arbitrary processor will succeed if multiple processes try to write to the same memory region at the same time.

2.4 LINKED PARTS IN MAPREDUCE AND ABOVE

Kiveris Raimondas [4] et al. As suggested in this system, a basic subroutine in graph clustering, computingrelated components of a graph is at the heart of many data mining techniques. Although this is a well- studied subject, many algorithms with strong theoretical guarantees turn out to be unsatisfactory in practice, especially when dealing with graphs that have billions or even hundreds of billions of edges. In this research, we develop enhanced methods for large-scale data analysis based on the conventional Map Reduce architecture. We also investigate the impact of adding a distributed hash table (DHT) service to Map Reduce. We demonstrate that these algorithms easily outperform previously investigated algorithms, often by more than an order of magnitude, and have verifiable theoretical guarantees. Specifically, our MapReduce implementation employing a DHT is 10 to 30 times faster than the best previously examined algorithms, and our iterative Map Reduce algorithms run 3 to 15 times faster than the best previously studiedalgorithms. Large-scale graph mining is an increasingly important subject in big data research and is a fundamental tool for modeling social, communication, and information networks. These are the fastest algorithms that scale to graphs with hundreds of billions of edges with ease. One step towards creating a general-purpose, user-friendly graph mining system is to do this computation in a popular, fault-tolerant distributed programming framework like Map Reduce or Hadoop. The essential first step that forms the basis of many more complex graph analysis algorithms is computing related components. The majority of previously researched algorithms that have strong theoretical guarantees may not work well in practice because they either (i) need an excessive number of calculation rounds (e.g., scaling with the graph's diameter). On a shared cluster with commodity technology, our techniques grow to graphs with billions ofnodes and hundreds of billions of edges with ease.

2.5 GRAPH CONNECTIVITY'S COMPLEXITY

Wigderson, Avi [5] et al. As suggested by this system, we review the key advancements in our knowledgeof the graph connection problem's complexity in a number of computational models in this study and point out a few difficult unsolved issues. You have to understand that graph connection is not a totally trivial problem if you have ever lost your way, even with a map and road signs to help you. Here (in theory), we are content to find out if there is a method at all, when in practice, the crucial problem is how to go from point A to point B. This is the computational problem that has been examined across the widest range of computing models, including decision trees, Boolean circuits, Turing machines, PRAMs, andcommunication complexity. It has shown to be a fruitful test case for contrasting fundamental resources, including time versus space, randomness versus determinism, and sequential versus parallel computing. The intricacy of network connectivity appears to be of great relevance for two complementary reasons. On the one hand, from a combinatorial perspective, it is straightforward enough to be virtually fully comprehended. However, its structure is sufficiently rich to capture (in various versions) a number of significant complexity classes, the precise amount of which is still unknown. Development in the area of connection complexity was patchy until approximately five years ago.

3. EXISTING SYSTEM

Numerous applications exist for determining connectedness in graphs, including neural networks, social network

research, data mining, and connectivity within or between cities. The vast array of graphapplications renders graph connectivity issues highly significant and merits more investigation. Many single-node graph mining and analysis methods are available at the moment, however they are mainly limited to tiny graphs and are implemented on a single computer node. Even with the most popular single-node algorithms, finding 2-edge connected components (2-ECCs) in enormous graphs (billions of edges and vertices) is computationally and practically unfeasible. Completing processing of a large graph in a distributed and parallel manner reduces processing time significantly. Furthermore, it makes stream data processing possible by providing speedy outcomes for large and continuous nature data sets. In order to detect 2-ECCs in large undirected graphs, this research suggests a distributed, parallel algorithm that it calls "BiECCA" and analyzes its time complexity. The suggested technique leverages an existing method to locate connected components (CCs) in a graph as a sub-step and is implemented on a Map Reduce framework. Finally, as a continuation of our work, we propose a few new concepts and methods.

4. PROPOSED SYSTEM

The integration of utility frequency pattern analysis and the power of Map Reduce and Directed Acyclic Graph (DAG), the suggested system is a sophisticated authentication framework designed to handle large amounts of data. The goal of this cutting-edge technology is to provide strong authentication procedures while addressing the difficulties associated with processing and analyzing big datasets. Scalability and parallel processing are made possible by the system's effective distribution and processing of authentication requests among a cluster of nodes, which makes use of the Map Reduce paradigm. One of the most important components in coordinating the authentication process is the Directed Acyclic Graph (DAG). Inorder to guarantee a seamless workflow free from cyclic dependencies, it arranges the authentication processes as a network of connected nodes. This architecture reduces the possibility of bottlenecks and optimizes resource allocation in addition to improving system reliability. The system uses Frequent Item set Utility Frequency Pattern Analysis to bolster security and identify questionable activity. Through repeated item set mining from authentication data, the system becomes proficient in recognizing possible risks or unauthorized access attempts by recognizing typical patterns and aberrant behaviors. By combining the power of Map Reduce, the Directed Acyclic Graph for smooth orchestration, and Frequent Item set Utility Frequency Pattern Analysis for improved security, this suggested system offers a novel method to large data authentication. In doing so, it gives businesses a strong weapon to protect their systems and data from unwanted access and potential cyber threats. It does this by offering a dependable, scalable, and secure authentication solution appropriate for contemporary, data-intensive situations.

4.1 BASE INFORMATION ANALYSIS

We can mine the entire set of frequent item sets in the base information analysis module, depending on howcomplete the patterns are that need to be mined. We can differentiate between the following types of frequent item set mining, given a minimum support threshold: the co-efficient, which denotes the variety of items, including the first or most significant item set. The item set is represented by the combitorial, andits length is denoted by the letter "j." When an item set has length 2 (j=2), it comprises both 1- and 2-item sets (i=1,2). The desired item set length is represented by the letter "m." m=k+1. Here, "m" stands for the length of the item set for which we will estimate the count. For example, if k=2 and m=3, then "k" denotesthe base information size. If k=2 in the base data, it indicates that there are two sets of items: one and two.

4.2 APPROMIZATION COUNT CALCULATION

The goal of this module is to produce the most frequent item sets with the least amount of work. This module adapts the idea of segmenting the data source and mining the segments for maximal frequent item sets, as opposed to creating candidates for determining maximal frequent item sets as done in previous approaches. It also minimizes the number of scans to only two over the transactional data source. Additionally, there is no longer any time needed for candidate generation. The following procedures are taken by this algorithm in order to ascertain from a data source:

4.3 FREQUENT ITEMSET LIST GENERATION

The sliding window model is applied in this module. There should be two sub-windows created from the sliding window. 'w' stands for the full window, and 'w0' and 'w1' stand for the sub-windows. The inputs should drive the dynamic partitioning of the sub-windows. From directed and undirected graph structured data with loops (including self-loops) and labeled or unlabeled nodes and links, it may generate all frequently occurring induced subgraphs. Utilizing applications for chemical carcinogenesis analysis and Web surfing pattern analysis, its performance is assessed in order to circumvent the issue of several database scans and the candidate generate-and-test procedure.

Directed Acyclic Graph (DAG) with Frequent Item set Utility Frequency Pattern Analysis Algorithm is the name of the corresponding algorithm. It simply takes two scans to get the database's information. Since the database's contents are stored in a tree form, frequent patterns are mined from it. In particular, the database is scanned once to identify all frequent 1- itemsets before the Incremental Directed Acyclic Graph (DAG) with Frequent Item set Utility Frequency

4.4 SKIP AND COMPLETE TECHNIQUE

This module divides the database into several non-overlapping pieces in order to generate a skip count. Locally frequent item sets in each segment can be determined following the initial database scanThis method thereby dramatically decreases the number of scans required by Apriori-based algorithms to just two. The partition technique is therefore constantly dependent on the number of segments and the distribution of the data. This counter is updated by deducting the appropriate "over-estimate" for each pattern item as the database is scanned. Any pattern including that item can be pruned if the counter falls below the minimal support. This is because the pattern cannot be repeated.

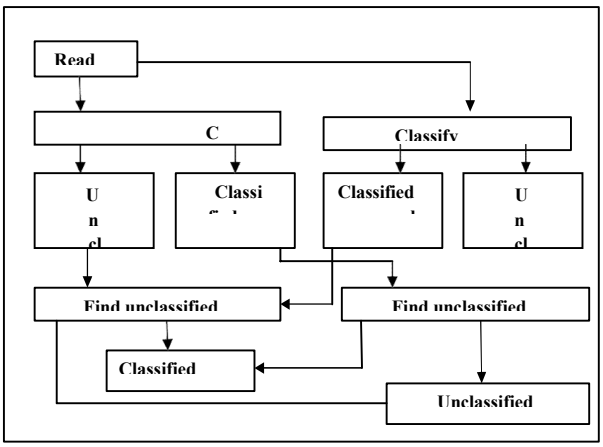

Fig 3. Block diagram

4.5 GROUP COUNT TECHNIQUE

The data report will be generated as a tree structure in this module. The algorithm attempts to decrease themining time by employing this structure. The process of the Incremental Directed Acyclic Graph (DAG) with Frequent Item set Utility Frequency Pattern Analysis algorithm only requires upkeep and updating ofthe multiple connections that connect one transaction containing a set of items to the next after the H-structhas been built. There is no need to scan the database more than once because Directed Acyclic Graph (DAG) with Frequent Item set Utility Frequency Pattern Analysis retains all transactions that contain frequent items in memory. After that, the H-struct is used to extract all of the data. In comparison to Directed Acyclic Graph (DAG) with Frequent Item set Utility Frequency Pattern Analysis, Incremental Directed Acyclic Graph (DAG) with Small Minimum Support Threshold performed better than Apriori in identifyingfrequent patterns more quickly and requiring less memory.

5. CONCLUSION

The suggested authentication system that combines utility frequency pattern analysis with Directed Acyclic Graphs (DAG) with frequent item sets offers a creative and promising solution to the problems associated with user authentication in the big data era. The system provides considerable performance and scalability benefits by using DAG-based computation to efficiently process large-scale authentication data. By spotting probable anomalies and reoccurring authentication patterns, the inclusion of frequent item set analysis improves security measures and offers insightful information for preventive security measures. The system is an attractive option for enterprises handling large volumes of authentication data because of its benefits,which include effective processing, scalability, improved security, and data-driven insights. The system can be successfully deployed and maintained, guaranteeing its dependability and robustness in real-world circumstances, with careful testing and strict implementation. The capacity of the proposed system to effectively authenticate users and evaluate authentication patterns becomes increasingly important for protecting sensitive data and preserving data integrity as the amount and complexity of

data continue to expand.

REFFERENCE

[1] alecologists. BioScience. 2002; 52: 19-30. 2. Yang S, Lho H-S and Song B. Sensor fusion for obstacle detection and its application to an unmanned ground vehicle. ICCAS-SICE, 2009. IEEE, 2009, p. 1365-9.

[2] YOUNG J, ELBANHAWI, E., and SIMIC, M. Developing a Navigation System for Mobile Robots. Intelligent Interactive Multimedia. Springer, 2015.

[3] Lowe DG. Distinctive image features from scale-invariant keypoints. International journal of computer vision. 2004; 60: 91-110.

[4] Ke Y and Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors. Computer Vision and Pattern Recognition, 2004 CVPR 2004 Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, p. II-506-II-13 Vol. 2.

[5] Al-Smadi, M., Abdulrahim, K., Salam, R.A. (2016). Traffic surveillance: A review of vision-based vehicle detection, recognition and tracking. International Journal of Applied Engineering Research, 11(1), 713–726

[6] C.Nagarajan and M.Madheswaran - 'Experimental verification and stability state space analysis of CLL-T Series Parallel Resonant Converter' - Journal of ELECTRICAL ENGINEERING, Vol.63 (6), pp.365-372, Dec.2012.

[7] C.Nagarajan and M.Madheswaran - 'Performance Analysis of LCL-T Resonant Converter with Fuzzy/PID Using State Space Analysis'- Springer, Electrical Engineering, Vol.93 (3), pp.167-178, September 2011.

[8] C.Nagarajan and M.Madheswaran - 'Stability Analysis of Series Parallel Resonant Converter with Fuzzy Logic Controller Using State Space Techniques'- Taylor & Francis, Electric Power Components and Systems, Vol.39 (8), pp.780-793, May 2011.

[9] C.Nagarajan and M.Madheswaran - 'Experimental Study and steady state stability analysis of CLL-T Series Parallel Resonant Converter with Fuzzy controller using State Space Analysis'- Iranian Journal of Electrical & Electronic Engineering, Vol.8 (3), pp.259-267, September 2012.

[10] Nagarajan C., Neelakrishnan G., Akila P., Fathima U., Sneha S. "Performance Analysis and Implementation of 89C51 Controller Based Solar Tracking System with Boost Converter" Journal of VLSI Design Tools & Technology. 2022; 12(2): 34–41p.

[11] C. Nagarajan, G.Neelakrishnan, R. Janani, S.Maithili, G. Ramya "Investigation on Fault Analysis for Power Transformers Using Adaptive Differential Relay" Asian Journal of Electrical Science, Vol.11 No.1, pp: 1-8, 2022.

[12] G.Neelakrishnan, K.Anandhakumar, A.Prathap, S.Prakash "Performance Estimation of cascaded h-bridge MLI for HEV using SVPWM" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:750-756

[13] G.Neelakrishnan, S.N.Pruthika, P.T.Shalini, S.Soniya, "Perfromance Investigation of T-Source Inverter fed with Solar Cell" Suraj Punj Journal for Multidisciplinary Research, 2021, Volume 11, Issue 4, pp:744-749

[14] C.Nagarajan and M.Madheswaran, "Analysis and Simulation of LCL Series Resonant Full Bridge Converter Using PWM Technique with Load Independent Operation" has been presented in ICTES'08, a IEEE / IET International Conference organized by M.G.R.University, Chennai.Vol.no.1, pp.190-195, Dec.2007

[15] M Suganthi, N Ramesh, "Treatment of water using natural zeolite as membrane filter", Journal of Environmental Protection and Ecology, Volume 23, Issue 2, pp: 520-530,2022

[16] M Suganthi, N Ramesh, CT Sivakumar, K Vidhya, "Physiochemical Analysis of Ground Water used for Domestic needs in the Area of Perundurai in Erode District", International Research Journal of Multidisciplinary Technovation, pp: 630-635, 2019

[17] Radhakrishnan, M. (2013). Video object extraction by using background subtraction techniques for sports applications. Digital Image Processing, 5(9), 91–97.

[18] Qiu-Lin, L.I., & Jia-Feng, H.E. (2011). Vehicles detection based on three-frame-difference method and cross-entropy threshold method. Computer Engineering, 37(4), 172–174.

[19] Liu, Y., Yao, L., Shi, Q., Ding, J. (2014). Optical flow based urban road vehicle tracking, In 2013 Ninth International Conference on Computational Intelligence and Security. https://doi.org/10.1109/cis.2013.89: IEEE

[20] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation, In 2014 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10. 1109/cvpr.2014.81: IEEE.

[21] Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M. (2013). Selective search for object recognition. International Journal of Computer Vision, 104(2), 154–171.

[22] Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 37(9), 1904–16

[23] Zhe, Z., Liang, D., Zhang, S., Huang, X., Hu, S. (2016). Traffic-sign detection and classification in the wild, In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) https://doi.org/10.1109/cvpr.2016.232: IEEE.

[24] Krause, J., Stark, M., Deng, J., Li, F.F. (2014). 3d object representations for fine-grained categorization, In 2013 IEEE International Conference on Computer Vision Workshops. https://doi.org/10.1109/iccvw.2013.77: IEEE.

[25] Yang, L., Ping, L., Chen, C.L., Tang, X. (2015). A large-scale car dataset for fine-grained categorization and verification, In 2015 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr. 2015.7299023 (pp. 3973–3981): IEEE.

[26] Zhen, D., Wu, Y., Pei, M., Jia, Y. (2015). Vehicle type classification using a semi supervised convolutional neural network. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2247–2256.

[27] Guerrero-Gomez-Olmedo, R., Torre-Jimenez, B., Lopez-Sastre, R., Maldonado-Bascon, S., Ooro-Rubio, D. (2015). Extremely overlapping vehicle counting, In Iberian Conference on Pattern Recognition & Image Analysis. https://doi.org/10.1007/978-3-319-19390-8_48 (pp. 423–431): Springer International Publishing