Plan for my Cooperative Work with INRIA and INRA in Montpellier during July 2012-July 2013

Dennis Shasha

May 2, 2011

My research over the last ten years has focussed largely on the application of computer science to the life sciences. Most of that work has had to do with systematic experimental design for biologists, visualization software, and machine learning tools. The current thrust of my work has to do with large scale inference and analysis of networks. My plan in Montpellier is to continue that work, thus melding with the Scalable Data Analysis theme of the Zenith and NUMEX work and leading to a tool that may be useful to the analysis of the data generated by the Integration of Nutritional Functions team (and possibly others) at INRA.

1 Brief Review of the Most Relevant of My Previous Work

My work in scientific computing is driven by the philosophy that the problems and questions should come from the lab scientists themselves. I attempt to solve those problems in a way that the scientist can use and to generalize that solution as much as possible. The following branches of my work all follow from that philosophy.

1. **Adaptive Combinatorial Design for the Design of Experiments** Lab scientists would like to answer a scientific question as quickly (in person time) and as economically (in lab equipment and materials) as possible. A typical "search space" in a lab setting will include many

possible perturbations (e.g. light, carbon, nitrogen, knock-outs,...) and the goal is to find the values of those purturbations that optimize a particular output (e.g. biomass, seed size). The expensive approach is to explore the entire search space. An alternative is to design a small number of experiments and then to use the results of those experiments to design the next group with a view towards finding the optimal conditions very rapidly. We have used and improved a technique from statistics called combinatorial design to this end[24]. Our basic strategy is to use combinatorial design (i) to design a well-spaced and very small set of initial experiments and (ii) to use the results of that first set of experiments to design a second set of experiments that focusses on the features that seem most influential. Thus, combinatorial design is used "recursively" to find the optimal values of the influential features and the other features. We have used this successfully in our plant biology group at NYU, but the method has been used by collaborators looking at bacteria and bioenergy.

2. **Visualization of Multiple Experiments**
   A frequent genomics question is "Which genes are most affected by all of these experiments or a subset of those experiments?" A common way to appreciate this visually is to use a Venn diagram. Because Venn diagrams generalize poorly beyond three experiments, we have developed a visual representation known as a Sungear (http://virtualplant.bio.nyu.edu/cgi-bin/sungear/index.cgi) that does generalize. Sungear is an interactive search interface that supports statistical conclusions and that is particularly strong in performing metaanalyses of many different experiments[22], [21]. For example, a sister lab is using Sungear for cancer studies. It is a general tool and we have seen applications ranging from science to marketing to the evaluation of sports teams.

3. **Data Analysis to infer gene or module function**
   Most of my work with Gloria Coruzzi, Ken Birnbaum, and Phil Benfey over the years has had to do with data analysis to discover gene function. Sometimes, this has meant the inference of the individual or combinatorial genetic causes of traits[26], sometimes in a cell-specific manner[25]. Most frequently though, the idea has been to take a holistic "systems biology" approach to try to understand the role of modules of genes, often using machine learning [23], [19], [18], [16], [13]. Many

of the tools we have developed are now incorporated into our system Virtual Plant[15].

4. **Network inference in genomic networks**
   The goal in this work is to determine which genes influence which other genes and the strength of those relationships. In the case of genomics, the experimental strategy consists of measuring the effect of perturbations (such as the introduction of stress, genetic change, or the insertion of nutrients) to organisms over time. The end result is a network that predicts causal relationships among genes. We have just begun that work [6, 4] and will outline our plan to continue it below.

5. **Subgraph queries**
   Related to the question of network inference is what to do when one has a large network or several large networks and one wants to find common motifs. The paradigmatic question is "where is a certain labeled query graph q in a large database D of graphs?" Our fundamental strategies have been to use filters to prune away graphs from D that cannot match q and then to use a location data structure to find good starting points for searches in the graphs of D that remain. We have explored several variants of this problem [7, 9, 20] Another question has to do with clustering graphs that are similar.[17]

6. **Time series analysis to find correlations and bursts among tens of thousands of time series over sliding windows**
   A paradigmatic problem in this area is to find highly correlated instruments in financial markets, where correlations can come and go over time. The two central techniques are to use dimensionality reduction techniques such as wavelets and sketches (random vectors) to avoid comparing all pairs of instruments and to update previous correlations efficiently. This work does not directly relate to our plant biology work, but could be used for other scientific applications having far longer time series [3, 11, 12].

2 The Basic Plan: Scalable Network Inference on a Workflow Platform

In our review of genomic network inference algorithms[29, 6, 32, 35, 38, 39], we have observed that there are several stages of analysis depending on the kind of data that is available. We can divide those data types along two dimensions: (i) whether that data is generated based on genetic perturbations or not; and (ii) whether that data consists of steady state data or time series data.

Example of genetic perturbations include the suppression of gene function or its enhancement, whether cell-specific of not. In a non-genomic setting, the equivalent of a genetic perturbation is any direct modification of a node in a network. Non-genetic perturbations, by contrast, are analogous to manipulations of the inputs of networks.

What constitutes time series data depends on the system under examination. For example, formally, measurements taken every 4 hours constitute a time series. For our purposes, however, time series data means a series of experiments having the property that the state at experimental time point k+1 depends on the state at time point k but not on the interaction among data components at k+1. For us, then, for a series of experiments to be considered a time series, causality edges flow from the state of elements (e.g., genes) at k to elements at k+1. Whether this is true for measurements taken every four hours or not will depend on the rate of reaction of elements in the organism under study. We might call the kind of time series we are interested in *causal time series*.

From steady state experiments, algorithms can derive correlation, clusters, and biclusters[31, 33, 37] Clustering reduces the number of nodes to a small group of "super-nodes" that rise and fall together. From genetic perturbations, one can determine the direct or indirect influence of a perturbed element on others.[29, 30, 39] From causal time series data, one can determine causal links – if there is enough data relative to the number of super-nodes.[34] The overall workflow can be represented in following Vistrails figure [36]

Whereas this offers a systematic way to infer networks, the resulting networks are not always so good. In our own work for example[6], we were able to predict the direction of gene expression (whether expression rose or fell) on out-of-sample data quite accurately, but not its magnitude. The main open

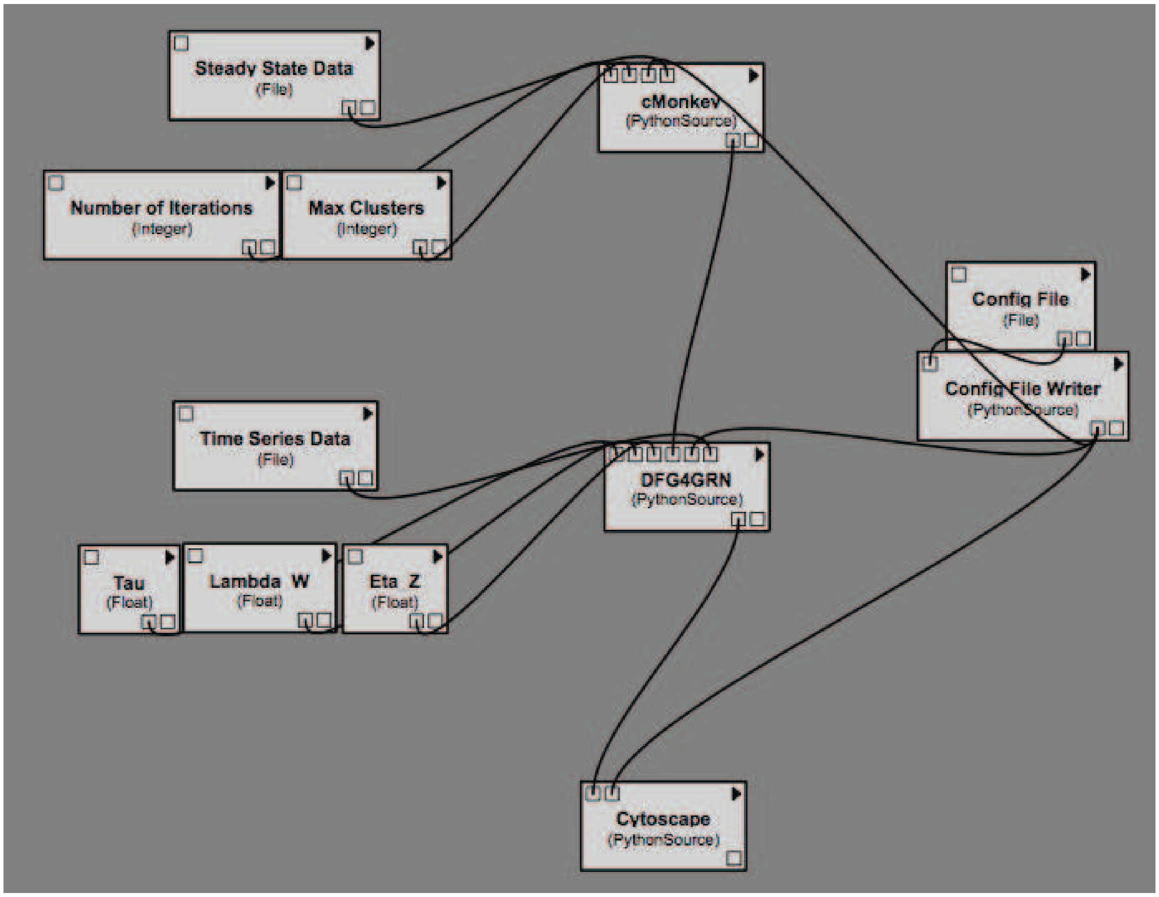

Figure 1: Steps of the network inference workflow: each step contains several optional algorithms and many parameters for each algorithm

problems in my view then are to (i) improve the quality of the algorithms given the data available and (ii) to determine which next experiment to do to improve the prediction accuracy.

2.1 Algorithms

There are two issues regarding algorithms: quality and speed. They are related, because faster algorithms make it possible to search more parameter combinations and thus achieve better quality.

The quality of a network inference workflow depends on the algorithms chosen and the parameters fed to those algorithms. For network inference the main algorithms are Inferelator 2.0[29, 34], ARACNE[40, 32], TSNI[35], BANJO[38], and NIR[39]. Different authors claim that each algorithm is best overall, but we suspect that each has a "sweet spot" which we must find. Finding the best parameters on the other hand will require an exploration of the parameter space. For this, we will use genetic algorithms[41, 42] in combination with combinatorial design. Here is where quality and speed interact: fast algorithms permit more exploration of the parameter space.

The speed issue comes up when the inference problem concerns large networks. The core problem concerns the identification of edges that could cause changes in the value of a target element and assigning values to them. This is essentially a regression problem. Luckily, there has recently been a flurry of excellent work on machine learning algorithms which are both sufficiently fast and parallelizable to be used on data sets with milions of elements. Below, we list a few of the algorithms we think might useful in training regression models on large-scale data.

- **Random Forests** [2]
  Random forests are ensembles of decision trees which are constructed from random subsets of the data. They're fast to train, easy to parallelize, and perform extremely well.

- **Large-Scale SVM Regression** [1]
  Bottou demonstrated that a stochastic gradient descent solver for a variety of learning problems (including support vector machine optimization) is able to scale with extremely large datasets while converging to the predictive performance of traditional optimization algorithms.

- **Large-Scale ℓ_1 Regularized Learning** [10]
  Stochastic coordinate descent can be used to learn sparse regression models, with small training times even for data sets where both the dimensionality and the number of training points is large.

2.2 Experiments to Do Next

Regardless of the quality of algorithms, insufficient or excessively noisy data can prevent an algorithm from inferring good networks. Because experiments take time and expense, we want to guide the experimenter to do the "right" experiment. One way to do this is to determine which existing experiment has been most valuable and doing another one like that. To determine the value of an existing experiment, one can remove that experiment, rerun the inference algorithm and then re-compute prediction accuracy.

For example, in the case of our Arabidopsis time-course study[6], removing two replicate experiments from two different timepoints prior to 15 minutes was less harmful to the accuracy of out-of-sample prediction of the network state at 20 min than removing both replicates from a single time-point prior to 15 min. This suggests that measurements at different time-points may be more valuable than replicates.

Whereas this rather naive approach may work well in some cases, it will not lead to radically different experimental designs. One way to discover better designs will be to simulate the data under different noise and variance assumptions. Given such simulation results, one should be able to approach a given application, characterize its noise and variance properties, and design a series of experiments.

2.3 The Importance of Workflow

Workflow systems help solve two important problems in scientific computation:

1. Just as it is important for experimental procedures to be repeatable so it is important for computational procedures to be repeatable. I have been active in urging computer scientists in the large database community to create repeatable experiments[47, 46] This year we featured the use of Vistrails to help make this possible (http://www.sigmod2011.org/calls_papers_sigmod_r

By storing workflows along with associated software and data, experimenters ensure that a whole computational flow can be reproduced.

2. Workflow systems support a disciplined appoach to parameter exploration. For example, Vistrails will soon have a genetic algorithms module so that parameter values can be varied and optimized.

For these reasons, our scalable network software will be wrapped in a workflow system, probably Vistrails to start.

3 Expected Results

Over the course of my sabbatical year (July, 2012 to July, 2013), I expect to design and build useful network inference software and apply it to important problems, I also intend to keep my eyes and ears open as new problems come to my attention. In my previous sabbaticals at INRIA Rocquencourt, I consistently found that interactions with colleagues led to new research directions and excellent publications[43, 44, 45]. I have every expectation that serendipity will play an equally positive role this time.

References

[1] L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] R. Cole, D. Shasha, and X. Zhao. Fast window correlations over uncooperative time series. In R. Grossman, R. J. Bayardo, and K. P. Bennett, editors, *KDD*, pages 743–749. ACM, 2005.

[4] R. A. Gutierrez, L. Lejay, A. Dean, F. Chiaromonte, D. E. Shasha, and G. M. Coruzzi. Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis, Jan. 2007.

[5] Z. Kimmel, D. Shasha, A. Turchin, and R. A. Greenes. State-Based Clinical Decision Support (SBCDS), June 2006.

[6] G. Krouk, P. Mirowski, Y. LeCun, D. Shasha, and G. Coruzzi. Predictive network modeling of the high-resolution dynamic plant transcriptome in response to nitrate. *Genome Biology*, 11(R123), December 2010.

[7] R. D. Natale, A. Ferro, R. Giugno, M. Mongiovì, A. Pulvirenti, and D. Shasha. SING: Subgraph search In Non-homogeneous Graphs, 2010.

[8] S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks, 2007.

[9] D. R. Recupero, R. A. Gutirrez, and D. Shasha. Graphclust: A Method for Clustering Database of Graphs.

[10] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l_1 regularized loss minimization. In A. P. Danyluk, L. Bottou, and M. L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, page 117. ACM, 2009.

[11] X. Zhao, X. Zhang, T. Neylon, and D. Shasha. Incremental methods for simple problems in time series: Algorithms and experiments, 2005.

[12] Y. Zhu and D. Shasha. *High performance discovery in time series: Techniques and case studies*, June 16 2003.

[13] Huang-Wen Chen, Sunayan Bandyopadhyay, Dennis E. Shasha, and Kenneth D. Birnbaum Estimation of genome-wide redundancy in Arabidopsis thaliana, BMC Evolutionary Biology 2010, 10:357; doi:10.1186/1471-2148-10-357

[14] X. Zhang, D. Shasha, Y. Song and J. T. L. Wang, Fast Elastic Peak Detection for Mass Spectrometry Data Mining, IEEE Transactions on Knowledge and Data Engineering, Issue 99. November 29, 2010, doi: 10.1109/TKDE.2010.238

[15] Manpreet S. Katari, Steve D. Nowicki, Felipe F. Aceituno, Damion Nero, Jonathan Kelfer, Lee Parnell Thompson, Juan M. Cabello, Rebecca S. Davidson, Arthur P. Goldberg, Dennis E. Shasha, Gloria M. Coruzzi, and Rodrigo A. Gutierrez, VirtualPlant: a software platform to support system biology research Plant Physiology 152:500-515 (2010)

[16] Gabriel Krouk, Daniel Tranchina, Laurence Lejay, Alexis A. Cruikshank, Dennis Shasha, Gloria M. Coruzzi, Rodrigo A. Guitierrez A Systems Approach Uncovers Restrictions for Signal Interactions Regulating Genome-wide Responses to Nutritional Cues in Arabidopsis PLOS Computational Biology March 2009, volume 5, issue 3

[17] Diego Reforgiato, Rodrigo Gutierrez, Dennis Shasha GraphClust: A Method for Clustering Databases of Graphs Journal of Information and Knowledge Management (JIKM) Volume: 7, Issue: 4 (December 2008) Page 231 - 241 http://www.worldscinet.com/cgi-bin/details.cgi?id=jsname:jikm&type=current

[18] Karen E Thum, Michael J Shin, Rodrigo Gutierrez, Indrani Mukherjee, Manpreet S Katari, Damion Nero, Dennis Shasha and Gloria M Coruzzi An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis" BMC Systems Biology 2008, 2:31 (04 Apr 2008)

[19] Gutierrez, R.A., Lejay, L., Chiaromonte, F., Shasha, D.E., Coruzzi, G.M. (2007) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol.: 8, pp. R7. "Must read" Factor 6 in the Faculty of 1000.

[20] A. Ferro, R. Giugno, M. Mongiovi, A. Pulvirenti, D. Skripin, D. Shasha, GraphFind: Enhancing Graph Searching by Low Support Data Mining Techniques BMC Bioinformatics, vol. 8 ISSN: 1471-2105, 2007.

[21] Rodrigo A. Gutirrez, Miriam L. Gifford, Chris Poultney, Rongchen Wang, Dennis E. Shasha, Gloria M. Coruzzi and Nigel M. Crawford Insights into the genomic nitrate response using genetics and the Sungear software system JXB Advance Access published online on April 29, 2007 Journal of Experimental Botany, doi:10.1093/jxb/erm079

[22] Christopher S. Poultney, Rodrigo A. Gutirrez, Manpreet S. Katari, Miriam L. Gifford, W. Bradford Paley, Gloria M. Coruzzi and Dennis E. Shasha Sungear: Interactive visualization and functional analysis of genomic datasets Bioinformatics, 2007; Jan 15;23(2):259-61 doi: 10.1093/bioinformatics/btl496

[23] Rodrigo Gutierrez, Dennis Shasha, and Gloria Coruzzi, Systems Biology for the Virtual Plant Plant Physiology, June 2005, vol. 38, pp. 550-554.

[24] Laurence V. Lejay, Dennis E. Shasha, Peter M. Palenchar, Andrei Y. Kouranov, Alexis A. Cruikshank, Michael F. Chou, Gloria M. Coruzzi Adaptive Combinatorial Design to explore Large Experimental Spaces: approach and validation Systems Biology, volume 1, issue 2, December 2004, pp. 206-212.

[25] Kenneth Birnbaum, Dennis E. Shasha, Jean Y. Wang, Jee W. Jung, Georgina M. Lambert, David W. Galbraith, and Philip N. Benfey A gene expression map of the Arabidopsis root Science, Dec 12 2003: 1956-1960 (A review article in the Research Focus section of Trends in Biotechnology called the article "At the end of 2003, the root biology community was blessed with what has become today already a historical paper that described for the first time a genome wide expression analysis of Arabidopsis root development [2].")

[26] Mitchell Levesque, Dennis Shasha, Wook Kim, Michael G. Surette, and Philip N. Benfey Trait-To-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes *Current Biology,* vol. 13, 129-133, January 21, 2003. Discussed in: http://www.the-scientist.com/yr2003/jun/hot_030603.html

[27] Alberto Lerner, Dennis Shasha, Zhihua Wang, Xiaojian Zhao, Yunyue Zhu Fast Algorithms for Time Series with Applications to Finance, Physics, Music, Biology and other Suspects ACM Sigmod 2004, pp. 965-968.

[28] Yunyue Zhu and Dennis Shasha Efficient Elastic Burst Detection in Data Streams The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD-2003 24 August 2003 - 27 August 2003

[29] Greenfield, A and Madar, A and Ostrer, H and Bonneau, R DREAM4: Combining Genetic and Dynamic Information to Identify Biological Networks and Dynamical Models PloS One 2010, vol. 5(10)

[30] Andrea Pinna and Nicola Soranzo and Alberto de La Fuente From Knockouts to Networks: Establishing Direct Cause-Effect Relationships through Graph Analysis PloS one, 5(10), e12912.

[31] Sepp Hochreiter and Ulrich Bodenhofer and Martin Heusel and Andreas Mayr and Andreas Mitterecker and Adetayo Kasim and Tatsiana Khamiakova and Suzy Van Sanden and Dan Lin and Willem Talloen and Luc Bijnens and Hinrich W. H. Ghlmann and Ziv Shkedy and Djork-Arn Clevert FABIA: Factor Analysis for Bicluster Acquisition Bioinformatics (2010).

[32] Adam A Margolin and Ilya Nemenman and Katia Basso and Chris Wiggins and Gustavo Stolovitzky and Riccardo Dalla Favera and Andrea Califano ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context BMC Bioinformatics 2006, 7(Suppl 1):S7

[33] David J Reiss and Nitin S Baliga and Richard Bonneau Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks BMC Bioinformatics 2006, 7:280

[34] Richard Bonneau and David Reiss and Paul Shannon and Marc Facciotti and Leroy Hood and Nitin Baliga and Vesteinn Thorsson The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo Genome Biology, 7(5), R36.

[35] Mukesh Bansal and Giusy Della Gatta and Diego Di Bernardo Inference of gene regulatory networks and compound mode of action from time course gene expression profiles Bioinformatics 2006, Vol. 22 no. 7, pages 815822

[36] Steven P. Callahan and Juliana Freire and Emanuele Santos and Carlos E. Scheidegger and Claudio T. Silva and Huy T. Vo VisTrails: Visualization meets Data Management In Proceedings of ACM SIGMOD 2006.

[37] Amos Tanay and Roded Sharan and Martin Kupiec and Ron Shamir Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data PNAS March 2, 2004 vol. 101 no. 9 2981-2986

[38] Jing Yu and V. Anne Smith and Paul P. Wang and Alexander J. Hartemink and Erich D. Jarvis Advances to Bayesian network inference for generating causal networks from observational biological data Bioinformatics 2004, vol. 20 issue 18

[39] Timothy S. Gardner and Diego di Bernardo and David Lorenz and James J. Collins Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling Science 4 July 2003: Vol. 301 no. 5629 pp. 102-105

[40] Pietro Zoppoli and Sandro Morganella and Michele Ceccarelli TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics 2010 Mar 25;11:154.

[41] Kalyanmoy Deb and Hans-Georg Beyer Self-adaptive genetic algorithms with simulated binary crossover Evol Comput. 2001 Summer;9(2):197-221.

[42] F. Herrera and M. Lozano and J.L. Verdegay Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis Artificial Intelligence Review, 12(4), 265-319.

[43] Dennis Shasha, F. Llirbat, E. Simon, P. Valduriez "Transaction Chopping: Algorithms and Performance Studies" *ACM Transactions on Database Systems*, October 1995, pp. 325-363.

[44] Francoise Fabret and Arno Jacobsen and Francois Llirbat and Jooa Pereira and Ken Ross and Dennis Shasha, "Filtering Algorithms and Implementation for Very Fast Publish/Subscribe Systems" SIGMOD, 2001, pp. 115-126.

[45] Nicolas Anciaux, Mehdi Benzine, Luc Bouganim, Philippe Pucheral, Dennis Shasha "GhostDB: querying visible and hidden data without leaks" SIGMOD Conference 2007: 677-688

[46] S. Manegold, I. Manolescu, L. Afanasiev, J. Feng, G. Gou, M. Hadjieleftheriou, S . Harizopoulos, P. Kalnis, K. Karanasos, D. Laurent, M. Lupu, N. Onose, C. Re, V . Sans, P. Senellart, T. Wu, and D. Shasha "Repeatability & Workability Evaluation of SIGMOD 2009" Sigmod Record, September 2009 http://www.sigmod.org/publications/sigmod-record/0909/index.html

[47] Ioana Manolescu, Loredana Afanasiev, Andrei Arion, Jens Dittrich, Stefan Manegold, Neoklis Polyzotis, Karl Schnaitter, Pierre Senellart, Spy-

ros Zoupanos, Dennis Shasha: "The repeatability experiment of SIGMOD 2008" SIGMOD Record 37(1): 39-45 (2008)