ΠΡΟΣ

1) **Ολα τα μέλη ΔΕΠ του Τμήματος Επιστήμης Υπολογιστών**
2) **Τους εκπροσώπους των Μεταπτυχιακών φοιτητών του Τμήματος Επιστήμης Υπολογιστών**
3) **Την Επταμελή Εξεταστική Επιτροπή**
4) **Ολα τα μέλη της Πανεπιστημιακής Κοινότητας**

Πρόσκληση σε Δημόσια Παρουσίαση της Διδακτορικής Διατριβής της

κα. Τριανταφύλλου Σοφία

Την Τετάρτη, 25 Φεβρουαρίου 2015 και ώρα 14:00 στην αίθουσα Κ206 τηλεδιάσκεψης του Τμήματος Επιστήμης Υπολογιστών του Πανεπιστημίου Κρήτης στο Ηράκλειο, θα γίνει η δημόσια παρουσίαση και υποστήριξη της Διδακτορικής Διατριβής της υποψήφιας διδάκτορας του Τμήματος Επιστήμης Υπολογιστών κα. Τριανταφύλλου Σοφία με θέμα:

"Ολοκληρωμένη αιτιακή ανάλυση ετερογενών συνόλων δεδομένων"

"Integrative causal analysis of heterogeneous data sets"

ΠΕΡΙΛΗΨΗ

Η επαναλαμβανόμενη μελέτη ενός συστήματος υπό διαφορετικές οπτικές για την εξαγωγή ενός συμπεράσματος είναι συχνό φαινόμενο στην επιστημονική πρακτική. Σε κάθε μελέτη, ο επιστήμονας συχνά μετρά διαφορετικές παραμέτρους του ίδιου συστήματος σε διαφορετικές πειραματικές συνθήκες. Το αποτέλεσμα μίας τέτοιας διαδικασίας είναι ένα σύνολο από ετερογενή σύνολα δεδομένων, που προέρχονται από διαφορετικές κατανομές. Κάθε σύνολο δεδομένων αναλύεται αυτοτελώς, και τα αποτελέσματα των αναλύσεων συντίθενται σε επιστημονική γνώση από την επιστημονική κοινότητα.

Παρ' όλη την ετερογένεια, σύνολα δεδομένων που μετρούν παραμέτρους του ίδιου συστήματος θα πρέπει να προέρχονται από, και άρα να αποτυπώνουν, τον ίδιο αιτιακό μηχανισμό. Υποστηρίζουμε ότι τέτοια σύνολα δεδομένων μπορούν να αναλυθούν μαζί βάσει αυτής της αρχής. Στη διατριβή αυτή, ορίζουμε και προτείνουμε μία λύση για το πρόβλημα του προσδιορισμού ενός ή όλων των πιθανών αιτιακών μηχανισμών που ταιριάζουν σε όλα

τα διαθέσιμα σύνολα δεδομένων ενός συστήματος. Ονομάζουμε αυτή την προσέγγιση ολοκληρωμένη αιτιακή ανάλυση.

Χρησιμοποιούμε τη γνωστή θεωρία της αιτιακής μοντελοποίησης, που συνδέει τις στατιστικές ιδιότητες ενός συνόλου δεδομένων με τον αιτιακό μηχανισμό που περιγράφει τις μετρούμενες μεταβλητές στο σύνολο αυτό. Πιο συγκεκριμένα, οι πολυπαραγοντικές σχέσεις των μετρούμενων μεταβλητών απότελούν περιορισμούς για τους πιθανούς αιτιακούς μηχανισμούς. Με αυτό τον τρόπο, το πρόβλημα μπορεί να διατυπωθεί σαν ένα πρόβλημα ικανοποίησης περιορισμών.

Η μέθοδος που προτείνουμε μεταφράζει τους στατιστικούς περιορισμούς που προκύπτουν από τα δεδομένα σε λογικές προτάσεις, μετατρέποντας το πρόβλημα εύρεσης πιθανού αιτιακού μηχανισμού σε ένα πρόβλημα ικανοποιησιμότητας (SAT). Περιορίζουμε την πολυπλοκότητα της μεθόδου με μία σειρά από ευριστικές ή ακριβείς βελτιώσεις. Εφόσον οι λογικές προτάσεις αντιστοιχούν σε στατιστικές σχέσεις, πιθανά αιτιακά σφάλματα οδηγούν σε μη ικανοποιήσιμες λογικές προτάσεις. Προτείνουμε μία μέθοδο για την αντιμετώπιση αυτού του προβλήματος που δεν επιβαρύνει την πολυπλοκότητα του αλγορίθμου. Τέλος, ταυτοποιούμε μία περίπτωση που η ολοκληρωμένη αιτιακή ανάλυση οδηγεί σε μία μη προφανή πρόβλεψη. Ελέγχουμε την ισχύ της πρόβλεψης αυτής σε μία ευρεία γκάμα δημόσιων δεδομένων, με στόχο να ελέγξουμε την επαληθευσιμότητα των υποθέσεων της αιτιακής μοντελοποίησης.

Δοκιμάσαμε τις μεθόδους μας σε μία πληθώρα διαφορετικών συνθηκών και συνόλων δεδομένων. Τα αποτελέσματα δείχνουν ότι (α) οι μέθοδοί μας έχουν την αναμενόμενη συμπεριφορά για διάφορες παραμέτρους εισόδου (β) οι μέθοδοί μας ξεπερνούν σε απόδοση τις σύγχρονες εναλλακτικές μεθόδους και (γ) αν και οι αιτιακές υποθέσεις δεν μπορούν να επαληθευτούν εύκολα, οδηγούν σε προβλέψεις που επαληθεύονται μαζικά σε πραγματικά σύνολα δεδομένων.

Επόπτης Διδακτορικής Διατριβής: Αναπλ. Καθηγητής, Ιωάννης Τσαμαρδίνος

ABSTRACT

Scientific practice typically involves repeatedly studying a system, each time trying to unravel a different perspective. In each study, the scientist may take measurements under different experimental conditions (interventions, manipulations, perturbations) and measure different sets of quantities (variables). The result is a collection of heterogeneous data sets coming from different data distributions. These data sets are analyzed in isolation and results are manually synthesized by the scientific community into scientific knowledge.

This thesis argues that heterogeneous data sets measuring the same system under study must all stem from, and therefore reflect, the same underlying causal mechanism, and that they can be co-analyzed based on this premise. We define the problem of identifying one or all

causal models that best fit all available data sets. We call this approach Integrative Causal Analysis.

The standard assumptions of causal modelling connect the statistical properties entailed in the available data sets to the underlying causal mechanism. Particularly, multivariate statistical relations of the measured variables constrain the search space of possible underlying causal models. Thus, the problem can be recast as a constraint satisfaction problem.

We propose an efficient conversion that translates statistical constraints into a SAT instance that can be solved with state-of-the-art SAT solvers. To improve scalability of our method we employ a series of approximate or exact steps that restrict the complexity of the conversion. Additionally, we introduce a scalable method for resolving conflicts arising from statistical errors. Finally, we identify a minimal example where INCA can produce a non-trivial prediction. We then test this prediction massively in public data sets from a wide range of scientific domains, in an attempt to test whether causally-inspired predictions are verified.

We test our methods in a variety of different data sets and conditions. Results indicate that (a) our methods are robust and behave reasonably against different input parameters (b) our methods outperform state-of-the-art alternatives and (c) while causal assumptions cannot be easily verified, they lead to statistical predictions that are massively validated in real-world data sets.

Supervisor: Accossiate Professor, Ioannis Tsamardinos

Panagiotis Tsakalides

Chairman

Department of Computer Science