ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διαχείριση Αβεβαιότητας στην Εξόρυξη Διεργασιών με τη χρήση Μπεϋζιανών Δικτύων

Θεοδωρόπουλος Νικόλαος-Παρασκευάς

Α.Μ. 71346393

Επιβλέπων: Αλέξανδρος Μπουσδέκης

Συν-επιβλέπων: Γεώργιος Μιαούλης

Περίληψη

Η παρούσα διπλωματική ερευνά την χρήση Μπεϋζιανών δικτύων για την διαχείριση της αβεβαιότητας στον τομέα της εξόρυξης διεργασιών. Αν και τα Μπεϋζιανά δίκτυα δεν είναι κάτι καινούργιο, δεν συνηθίζεται η εφαρμογή τους στην εξόρυξη διεργασιών. Αφού πρώτα μελετήσουμε κάποιους αλγορίθμους που χρησιμοποιούνται στην εξόρυξη διεργασιών για την ανακάλυψη διαδικασιών στη συνέχεια θα προσπαθήσουμε να συνδυάσουμε τους δύο αυτούς τομείς, των Μπεϋζιανών δικτύων και της εξόρυξης διεργασιών δηλαδή, έτσι ώστε να ερευνήσουμε αν γίνεται να προβλεφθούν οι καταστάσεις των δραστηριοτήτων που αποτελούν μια διαδικασία. Τέλος θα παρουσιάσουμε τα αποτελέσματα των προβλέψεων που θα γίνουν χρησιμοποιώντας ένα Μπεϋζιανό δίκτυο που θα έχει κατασκευαστεί από ένα σύνολο δεδομένων μιας διαδικασίας.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Εξόρυξη διεργασιών, μάθηση δομής, Μπεϋζιανά δίκτυα, διαχείριση επιχειρησιακών διαδικασιών

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η Θεοδωρόπουλος Νικόλαος-Παρασκευάς του Γεωργίου, με αριθμό μητρώου 71346393 φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι: «Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα

Georgios Miaoulis
Digitally signed by Georgios Miaoulis
DN: c=GR, l=Athens, o=University of West Attica, ou=Class B - Private Key created and stored in software CSP, sn=Miaoulis, givenName=Georgios, serialNumber=3179983822, cn=Georgios Miaoulis, email=gmiaoul@uniwa.gr
Date: 2021.07.25 16:37:23 +03'00'

PARIS MASTOROKOSTAS
Digitally signed by PARIS MASTOROKOSTAS
Date: 2021.07.25 17:12:07 +03'00'

Athanasios Voulodimos
Digitally signed by Athanasios Voulodimos
Date: 2021.07.25 23:17:22 +03'00'

Περιεχόμενα

Πίνακας εικόνων

1 Διαχείριση Επιχειρησιακών Διαδικασιών – Business Process Management (BPM)

Στο πρώτο αυτό κεφάλαιο θα αναφερθούμε στην διαχείριση επιχειρησιακών διαδικασιών. Αρχικά θα ορίσουμε τι είναι η διαχείριση επιχειρησιακών διαδικασιών, ενώ στη συνέχεια θα κάνουμε μια ιστορική αναδρομή της και στο τέλος θα παρουσιάσουμε το κύκλο ζωής της διαχείρισης επιχειρησιακών διαδικασιών. Ο στόχος του κεφαλαίου είναι να κατανοήσουμε τον τομέα πάνω στον οποίο θέλουμε να δοκιμάσουμε τη χρήση των Bayesian Network για την εξόρυξη διεργασιών.

1.1 Ορισμός

Η διαχείριση επιχειρηματικών διαδικασιών (BPM) περιλαμβάνει μεθόδους, τεχνικές και εργαλεία για την υποστήριξη του σχεδιασμού, της εφαρμογής, της διαχείρισης και της ανάλυσης τέτοιων επιχειρησιακών διαδικασιών [1]. Τα τελευταία χρόνια έχει τραβήξει την προσοχή λόγω της σημαντικής αύξησης παραγωγικότητας και ταυτόχρονα μείωσης των εξόδων που μπορεί να προσφέρει.

Η BPM μπορεί να θεωρηθεί ως επέκταση των κλασικών συστημάτων και προσεγγίσεων διαχείρισης ροής εργασιών (Workflow Management – WFM). Η διαχείρισης ροής εργασιών εστιάζει στην αυτοματοποίηση των επιχειρησιακών διαδικασιών, ενώ η διαχείριση επιχειρηματικών διαδικασιών καλύπτει ένα ευρύτερο φάσμα, από την αυτοματοποίηση και την ανάλυση διαδικασιών μέχρι την διαχείριση διαδικασιών και την οργάνωση της εργασίας. Η BPM στοχεύει στην βελτίωση επιχειρησιακών διαδικασιών, όσο το δυνατόν χωρίς την χρήση νέων τεχνολογιών. Για παράδειγμα μια εταιρεία μπορεί να κατανοήσει με ποιους τρόπους μπορεί να μειώσει τα κόστη ενώ ταυτόχρονα βελτιώνει το επίπεδο των υπηρεσιών της δημιουργώντας μοντέλα των επιχειρησιακών διαδικασιών και αναλύοντάς τα μέσω προσομοιώσεων. Από την άλλη η διαχείριση επιχειρηματικών διαδικασιών συχνά συνδέεται με λογισμικό που διαχειρίζεται, ελέγχει και υποστηρίζει επιχειρησιακές διαδικασίες. Σε αυτό εστίαζε αρχικά η διαχείριση ροής εργασιών, όμως οι κλασικές τεχνολογίες στόχευαν στην αυτοματοποίηση των διαδικασιών με έναν μηχανικό τρόπο χωρίς να δίνει την απαραίτητη προσοχή στον ανθρώπινο παράγοντα.

1.2 Ιστορική αναδρομή

Για να αναδειχθεί η σημασία των συστημάτων διαχείρισης επιχειρηματικών διαδικασιών μπορούμε να τα παρουσιάσουμε μέσω μιας ιστορικής επισκόπησης. Οι ρίζες της διαχείρισης επιχειρηματικών διαδικασιών εκτείνονται από την επιστήμη των υπολογιστών μέχρι την διοικητική επιστήμη. Επομένως είναι δύσκολο να προσδιοριστεί ακριβώς η αρχή της BPM. Από την βιομηχανική επανάσταση η παραγωγικότητα των επιχειρήσεων αυξάνεται συνεχώς λόγω τεχνικών καινοτομιών, βελτιώσεων στην οργάνωση της δουλειάς και της χρήσης της πληροφορικής. Την σήμερον ημέρα αν παρομοιάσουμε ένα πληροφοριακό σύστημα με μια σφαίρα, αυτό θα αποτελείται από ένα πλήθος στρώσεων. Στον πυρήνα που αποτελεί τη πρώτη στρώση αυτής της σφαίρας βρίσκεται το λειτουργικό σύστημα, δηλαδή το λογισμικό που κάνει το υλικό (hardware) να δουλεύει. Στην συνέχεια έχουμε τις εφαρμογές γενικού χαρακτήρα, τις οποίες μπορούν να χρησιμοποιηθούν από πολλές διαφορετικές επιχειρήσεις. Επιπλέον αυτές οι εφαρμογές συνήθως χρησιμοποιούνται από πολλαπλά τμήματα της ίδιας εταιρείας. Παραδείγματα τέτοιων εφαρμογών είναι ένα σύστημα βάσης δεδομένων, ένας συγκεκριμένος γραφέας κειμένου, το excel κοκ. Το τρίτο επίπεδο αποτελείται από εφαρμογές που στοχεύουν σε συγκεκριμένους τομείς, οι οποίες χρησιμοποιούνται από συγκεκριμένες επιχειρήσεις και τμήματα. Για παράδειγμα ένα σύστημα παρακολούθησης ασθενών για τα νοσοκομεία ή ένα σύστημα GPS για τα αμάξια. Στο τελευταίο επίπεδο βρίσκονται εξατομικευμένες εφαρμογές, δηλαδή εφαρμογές ανεπτυγμένες για ένα πολύ συγκεκριμένο οργανισμό ή επιχείρηση. Για παράδειγμα ένα σύστημα παρακολούθησης της παραγωγής για ένα συγκεκριμένο εργοστάσιο.

Την δεκαετία του 60’ το δεύτερο και το τρίτο επίπεδο δεν υπήρχαν, καθώς τα πληροφοριακά συστήματα ήταν ανεπτυγμένα πάνω σε μικρά λειτουργικά συστήματα με περιορισμένες δυνατότητες. Εφόσον κανένα λογισμικό γενικού χαρακτήρα ή λογισμικό με στόχο κάποιο συγκεκριμένο πεδίο δεν υπήρχε τα περισσότερα πληροφοριακά συστήματα φτιάχνονταν εξατομικευμένα για κάθε οργανισμό.

Περίπου το 1950 οι υπολογιστές και οι ψηφιακές επικοινωνίες ξεκίνησαν να επηρεάζουν την διαχείριση των διαδικασιών. Αυτό είχε ως αποτέλεσμα την δραματική αλλαγή στην οργάνωση των εργασιών και επέτρεψε νέους τρόπους επιχειρησιακής δραστηριότητας. Σήμερα οι καινοτομίας στον χώρο των υπολογιστών και των επικοινωνιών είναι ακόμα οι βασικοί άξονες που ωθούν σε αλλαγές στον τρόπο διαχείρισης διαδικασιών. Πλέον οι επιχειρησιακές

διαδικασίες είναι πιο περίπλοκες, βασίζονται σε πληροφοριακά συστήματα και μπορεί να εκτείνονται σε πολλούς οργανισμούς. Επομένως γίνεται εύκολα κατανοητό πως η μοντελοποίηση των διαδικασιών είναι ύψιστης σημασίας. Τα μοντέλα διαδικασιών βοηθούν στην διαχείριση της πολυπλοκότητας, παρέχοντας γνώση και τεκμηρίωση των διεργασιών. Η εξέλιξη όμως των τελευταίων χρόνων έχει αυξήσει τις δυνατότητες όλων των στρώσεων. Αποτέλεσμα αυτής της εξέλιξης η μετατόπιση της προσοχής από την ανάπτυξη λογισμικού στην συναρμολόγηση πολύπλοκων συστημάτων με έτοιμο λογισμικό. Πλέον η πρόκληση δεν είναι στην δημιουργία νέου κώδικα αλλά στην ένωση κώδικα και από τις τέσσερις στρώσεις.

Μια άλλη μετατόπιση είναι από τα δεδομένα στις διαδικασίες. Στις δεκαετίες του 80’ και του 90’ κυριάρχησαν οι προσεγγίσεις που βασίζονταν σε δεδομένα. Στο επίκεντρο της πληροφορικής ήταν η αποθήκευση και η ανάκτηση πληροφοριών και ως αποτέλεσμα η μοντελοποίηση των δεδομένων ήταν η βάση για το χτίσιμο ενός πληροφοριακού συστήματος. Επιπλέον νέες τάσεις στην διοίκηση των επιχειρήσεων όπως ο επανασχεδιασμός επιχειρηματικών διαδικασιών δείχνει την αυξημένη έμφαση στις διαδικασίες και όχι στα δεδομένα. Ως αποτέλεσμα τα νέα συστήματα βασίζονται σε προσεγγίσεις που δίνουν έμφαση στις διαδικασίες.

Τέλος θα πρέπει να αναφέρουμε μια ακόμα αλλαγή στον τρόπο που προσεγγίζονται τα πληροφοριακά συστήματα. Από συστήματα που ήταν πολύ προσεκτικά σχεδιασμένα σε συστήματα που συνεχώς επανασχεδιάζονται και μεγαλώνουν οργανικά. Δηλαδή πλέον τα συστήματα δεν μένουν στον αρχικό σχεδιασμό και αλλάζουν συνεχώς με βάση νέες τεχνολογίες που μπορεί να προκύψουν ή λόγω νέων προκλήσεων και απαιτήσεων που δημιουργούνται. Βασικός λόγος αυτής της ανάγκης για συνεχής αλλαγή είναι φυσικά το διαδίκτυο και ο τρόπος με τον οποίο λειτουργεί και συνεχώς εξελίσσεται. Αντίκτυπος των παραπάνω είναι πολύ λιγότερα συστήματα να φτιάχνονται πλέον από το μηδέν, αντίθετα η προσέγγιση που επιλέγεται είναι η σύνθεση έτοιμων εφαρμογών.

1.3 Κύκλος Ζωής της Διαχείρισης Επιχειρηματικών Διαδικασιών – BPM life-cycle

Υπάρχουν πολλές διαφορετικές εκδόσεις του κύκλου ζωής της διαχείρισης επιχειρηματικών διαδικασιών, όμως η βάση παραμένει η ίδια. Μπορεί να ειπωθεί πως ο κύκλος ζωής αποτελείται από τέσσερεις φάσεις [3]. Όπως φαίνεται και στην παρακάτω εικόνα (Εικόνα 1) οι τέσσερεις

φάσεις του κύκλου ζωής της διαχείρισης επιχειρηματικών διαδικασιών συνθέτουν μια επανάληψη με σκοπό την συνεχής βελτίωση των διαδικασιών μιας επιχείρησης.

Εικόνα 1: Ο κύκλος ζωής της διαχείρισης επιχειρηματικών διαδικασιών

- **Σχεδιασμός διαδικασίας**

  Κάθε προσπάθεια για διαχείριση επιχειρησιακών διαδικασιών απαιτεί την μοντελοποίηση μιας υπάρχουσας ή επιθυμητής διαδικασίας. Κατά την διάρκεια αυτής της φάσης κατασκευάζονται μοντέλα διαδικασιών που περιλαμβάνουν διάφορες οπτικές (ροή-ελέγχου, ροή-δεδομένων, οργανωτική, κοινωνική ,εννοώντας σχέσεις μεταξύ των εμπλεκόμενων μερών, και υπηρεσιακή). Ο μόνος τρόπος να δημιουργηθεί ένα «process-aware» επιχειρησιακό πληροφοριακό σύστημα είναι η προσθήκη γνώσης για την παρούσα επιχειρησιακή διαδικασία.
- **Διαμόρφωση συστήματος**

  Με βάση τη προηγούμενη φάση το «process-aware» επιχειρησιακό πληροφοριακό σύστημα θα πάρει σάρκα και οστά. Παραδοσιακά η πραγματοποίηση ενός συστήματος

θα απαιτούσε μια πολύπλοκη και χρονοβόρα διαδικασία ανάπτυξης λογισμικού. Για τον λόγο αυτό χρησιμοποιούνται έτοιμα λογισμικά που αντικαθιστούν την χρονοβόρα διαδικασία της ανάπτυξης νέου λογισμικού με μια φάση διαμόρφωσης και συναρμολόγησης διαδικασιών. Επομένως χρησιμοποιούμε τον όρο «διαμόρφωση συστήματος» για την φάση ανάμεσα στον σχεδιασμό και την εφαρμογή.

- **Εφαρμογή διαδικασίας**

  Το επόμενο βήμα είναι η εφαρμογή του «process-aware» επιχειρησιακού πληροφοριακού συστήματος που σχεδιάστηκε και διαμορφώθηκε στις προηγούμενες φάσεις. Όπως γίνεται εύκολα κατανοητό όσο διεξοδικά και αν γίνει ο σχεδιασμός των διαδικασιών και η διαμόρφωση του συστήματος τα πάντα κρίνονται στη εφαρμογή του μοντέλου σε πραγματικές συνθήκες. Μόνο έτσι μπορούμε στο επόμενο βήμα να αξιολογήσουμε το αποτέλεσμα καθώς και σημεία που χρήζουν βελτίωσης.

- **Διάγνωση**

  Κάθε «process-aware» επιχειρησιακό πληροφοριακό σύστημα θα πρέπει στην πάροδο του χρόνου να βελτιώνει τις επιδόσεις του, αξιοποιεί νέες τεχνολογίες, υποστηρίζει νέες διαδικασίες και να προσαρμόζεται σε ένα συνεχώς μεταβαλλόμενο περιβάλλον. Αυτός είναι και ο λόγος για τον οποίο το σύστημα που ξεκίνησε να εκτελείται από τη προηγούμενη φάση πρέπει να αξιολογείται, να υπάρχει μια διεξοδική διάγνωση έτσι ώστε να υπάρχει επαρκής ανατροφοδότηση (feedback) για να ξεκινήσει ξανά ο κύκλος ζωής της διαχείρισης επιχειρησιακών διαδικασιών με απώτερο την ασταμάτητη βελτίωση των συστημάτων μιας επιχείρησης.

2 Εξόρυξη Διεργασιών – Process Mining

Αφού πλέον κατανοήσαμε τα βασικά μέρη της διαχείρισης επιχειρησιακών διαδικασιών τώρα σε αυτό το κεφάλαιο θα αναφερθούμε στην εξόρυξη διεργασιών. Μέσω της εξόρυξης διεργασιών στοχεύουμε στην κατανόηση, ανάλυση και βελτιστοποίηση των επιχειρησιακών διαδικασιών. Αρχικά παρουσιάζουμε τα βασικά στοιχεία της εξόρυξης διεργασιών, όπου θα δοθεί και ένας απλός ορισμός. Στη συνέχεια γίνεται μια σύντομη παρουσίαση των αρχείων καταγραφής δεδομένων, πάνω στο οποίο στηρίζεται η εξόρυξη διεργασιών. Τέλος θα γίνει αναφορά στις λειτουργίες/μορφές της εξόρυξης διεργασιών.

2.1 Βασικά στοιχεία

Η εξόρυξη διεργασιών είναι ένα σχετικά καινούργιο ερευνητικό πεδίο το οποίο προσπαθεί να ανακαλύψει, να παρακολουθήσει και να βελτιώσει τις διαδικασίες μέσω της εξαγωγής γνώσης από ένα σύνολο γεγονότων [2]. Ως διαδικασία ορίζεται ένα σύνολο ενεργειών με καθορισμένη σειρά που εκπληρώνουν ένα σκοπό. Για παράδειγμα, όλες τα βήματα που πρέπει να ακολουθήσει ένας πολίτης για την έκδοση ενός δανείου. Οι διαδικασίες δεν περιορίζονται μόνο στο κομμάτι των επιχειρήσεων, διαδικασία μπορεί να θεωρηθεί και ο τρόπος με τον οποίο ένας άνθρωπος ετοιμάζεται κάθε πρωί για να πάει στη δουλειά του.

Η εξόρυξη διαδικασιών περιλαμβάνει την ανακάλυψη διαδικασιών (δηλαδή την εξαγωγή μοντέλων μιας διαδικασίας από αρχεία καταγραφής γεγονότων (event logs)), τον έλεγχο συμμόρφωσης (δηλαδή τον εντοπισμό και την παρακολούθηση των αποκλίσεων μεταξύ πρότυπου μοντέλου διαδικασίας και ιστορικού γεγονότων (conformance checking)), την εξαγωγή διαφόρων δικτύων (όπως κοινωνικών δικτύων και δικτύων οργανισμού), την αυτόματη δημιουργία και προσομοίωση μοντέλων, την επέκταση και επιδιόρθωση αυτών, καθώς και την πρόγνωση περιστατικών βάσει ιστορικών στοιχείων.

Σκοπός της εξόρυξης διεργασιών είναι η διεύρυνση των υπαρχουσών προσεγγίσεων που επικεντρώνονται στην βελτίωση της ευφυίας των επιχειρήσεων, όπως για παράδειγμα η βελτίωση επιχειρηματικών διαδικασιών (Business Process Improvement), η διοίκηση ολικής ποιότητας (Total Quality Management) και το Six Sigma. Τα τελευταία χρόνια παρατηρείται μια ωρίμαση των τεχνικών που χρησιμοποιούνται στην εξόρυξη διεργασιών, ενώ παράλληλα τα διαθέσιμα δεδομένα που αφορούν αρχεία καταγραφής γεγονότων είναι ευρέως διαθέσιμα χάρη

στην ψηφιοποίηση των επιχειρηματικών διαδικασιών και την σχεδόν καθολική χρήση πληροφοριακών συστημάτων από τις επιχειρήσεις.

Παραδοσιακά η προσπάθεια για την βελτίωση των διεργασιών σε έναν οργανισμό βασιζόταν στην δημιουργία και στην ανάλυση μοντέλων σε θεωρητικό επίπεδο, που όμως σπανίως ανταποκρίνονταν στην πραγματικότητα. Στόχος της εξόρυξης διεργασιών είναι να συμπληρώσει το κενό που υπήρχε μεταξύ των θεωρητικών μοντέλων και των πραγματικών γεγονότων. Δεν θα πρέπει να μπερδεύουμε την εξόρυξη διεργασιών με την εξόρυξη δεδομένων. Οι ρίζες της εξόρυξης διεργασιών βρίσκονται στην διαχείριση επιχειρηματικών διαδικασιών.

2.2 Αρχεία Καταγραφής Γεγονότων - Event Logs

Η εξόρυξη διεργασιών εκμεταλλεύεται την ανάλυση των καταγεγραμμένων γεγονότων κατά την εκτέλεση των επιχειρηματικών (ή μη) δραστηριοτήτων. Πιο πάνω ορίσαμε την διαδικασία ως ένα σύνολο βημάτων, τα βήματα αυτά μπορούν να θεωρηθούν ως γεγονότα. Για παράδειγμα, η υποβολή μιας αίτησης έκδοσης δανείου μπορεί να είναι το πρώτο βήμα για την διαδικασία της έκδοσής του αλλά και ένα γεγονός, ότι δηλαδή κάποιος, κάποια συγκεκριμένη στιγμή υπέβαλε της συγκεκριμένη αίτηση.

Σημείο εκκίνησης της εξόρυξης διεργασιών είναι η εγγραφή – καταχώρηση ενός γεγονότος (event log). Ένα αρχείο καταγραφής είναι ένα αρχείο στο οποίο καταγράφονται γεγονότα ή δεδομένα που λαμβάνουν χώρα σε ένα λειτουργικό σύστημα, κατά την εκτέλεση κάποιου λογισμικού ή κατά την επικοινωνία χρηστών. Βασική παραδοχή των τεχνικών εξόρυξης διεργασιών είναι πως υπάρχει η δυνατότητα να καταγραφεί μια αλληλουχία γεγονότων, κάθε ένα από τα οποία αφορά μια εργασία, δηλαδή ένα σαφώς προσδιορισμένο βήμα μιας διαδικασίας και συνδέεται με μια συγκεκριμένη περίπτωση εκτέλεσης ,δηλαδή με ένα στιγμιότυπο διαδικασίας. Τα πιο σημαντικά στοιχεία ενός γεγονότος είναι το είδος της εργασίας (Activity), ένα μοναδικό αναγνωριστικό της συγκεκριμένης εκτέλεσης (Case ID), και ο χρόνος εκτέλεσης της εργασίας (Timestamp).

Η ευρύτερη διαδικασία κατασκευής αρχείων καταγραφής γεγονότων από τα ακατέργαστα δεδομένα, απαιτεί 4 προϋποθέσεις:

- Επιλογή των γεγονότων που είναι σχετικά με την προς εξέταση διεργασία

- Έλεγχος της μεταξύ τους συσχέτισης προκειμένου να αποτελούν μία εκτέλεση της διεργασίας (ένα ίχνος)
- Ταξινόμηση γεγονότων βάσει της χρονοσφραγίδας εκτέλεσής τους
- Συμπλήρωση των χαρακτηριστικών των γεγονότων από τα αρχικά δεδομένα

Με βάση τα παραπάνω ορίζονται οι κατευθύνσεις-περιορισμοί προκειμένου το εξαγόμενο event log να αποτελεί ένα καλό σημείο αναφοράς για την Εξόρυξη Διεργασιών:

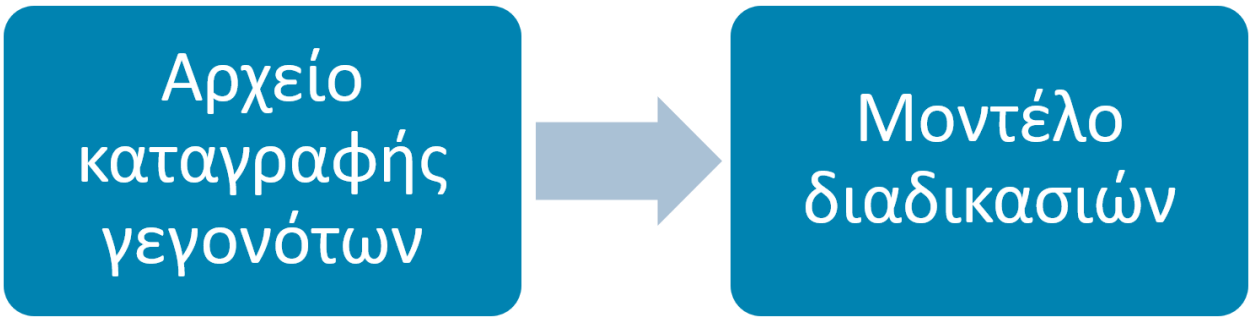
- Τα ονόματα των αναφορών και των χαρακτηριστικών των γεγονότων πρέπει να έχουν ξεκάθαρη σημασία, η οποία γίνεται αντιληπτή από όλους με τον ίδιο τρόπο.
- Τα ονόματα των αναφορών και των χαρακτηριστικών πρέπει να προέρχονται από μία δομημένη και διαχειρίσιμη συλλογή, με τη μορφή μιας ταξινομίας ή οντολογίας.
- Οι αναφορές πρέπει να είναι σταθερές. Για παράδειγμα να μην επαναχρησιμοποιούνται τα μοναδικά αναγνωριστικά τους και να μη βασίζονται σε μεταβλητές παραμέτρους.
- Οι τιμές των χαρακτηριστικών πρέπει να είναι όσο περισσότερο ακριβείς είναι εφικτό. Για παράδειγμα η χρονοσφραγίδα των γεγονότων να ακολουθεί το ίδιο επίπεδο ακρίβειας σε κάθε καταγραφή, δηλαδή να συμπληρώνεται και η ώρα και όχι μόνο η ημερομηνία.

2.3 Μορφές της εξόρυξης διεργασιών

Στο κεφάλαιο αυτό γίνεται μια προσπάθεια να κατηγοριοποιηθούν οι τεχνικές της εξόρυξης δεδομένων σύμφωνα με τις λειτουργίες που εκτελούν. Οι κατηγορίες αυτές είναι τέσσερις και είναι αρχικά η ανακάλυψη των διεργασιών (process discovery), στη συνέχεια ο έλεγχος συμμόρφωσης (conformance checking), ο επανασχεδιασμός διαδικασιών (process reengineering) καθώς και η επιχειρησιακή υποστήριξη (operational support) [5].

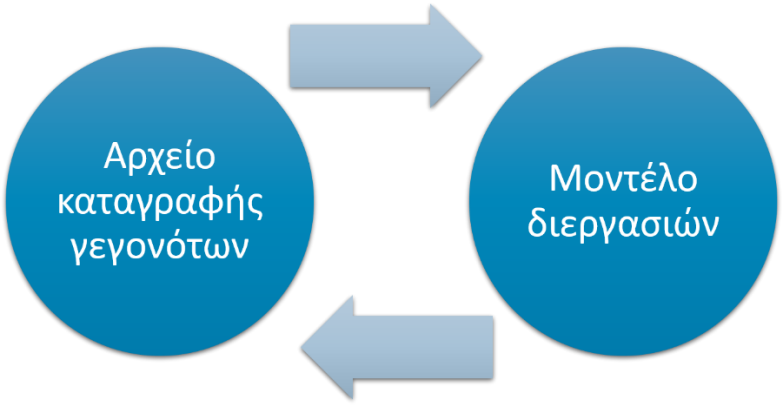
Στην ανακάλυψη διεργασιών ανήκουν το σύνολο των τεχνικών που εξάγουν μοντέλα διεργασιών χρησιμοποιώντας αποκλειστικά τα δεδομένα από καταχωρημένα γεγονότα. Στα αρχεία καταγραφής γεγονότων υπάρχει πληθώρα δεδομένων, στόχος της ανακάλυψης διεργασιών είναι να βρει τις σχέσεις που συνδέουν αυτά τα δεδομένα, ποια από αυτά δείχνουν τον τρόπο που εκτελείται μια διαδικασία και ποιες ενέργειες αποτελούν κάθε διαδικασία που ανακαλύπτεται. Από την ανάλυση αυτών των δεδομένων μπορούν να βγουν χρήσιμα συμπεράσματα για τις επιχειρηματικές διαδικασίες των εκάστοτε οργανισμών. Στην παρακάτω εικόνα (εικόνα 3)

βλέπουμε την σχέση μεταξύ του μοντέλου διαδικασιών και του αρχείου καταγραφής, καθώς και πως από το αρχείο καταλείγουμε στο μοντέλο.



Εικόνα 2 : Ανακάλυψη διεργασιών

Έλεγχος συμμόρφωσης καλείται η λειτουργία κατά την οποίο συγκρίνεται ένα θεωρητικό μοντέλο διαδικασιών με τα καταγεγραμμένα γεγονότα από την εκτέλεσή της. Έτσι γίνεται μια ανίχνευση και διάγνωση τόσο των διαφορών όσο και των κοινών σημείων ανάμεσα στο θεωρητικό μοντέλο και την πραγματικότητα. Αυτή η σύγκριση, όπως φαίνεται και στην εικόνα 3, είναι ένας τρόπος να αξιολογηθεί κατά πόσο μοντέλα διεργασιών που δημιουργούνται ανταποκρίνονται στη πραγματικότητα.



Εικόνα 3 : Έλεγχος συμμόρφωσης

Ο επανασχεδιασμός προσπαθεί να βελτιώσει ή να επεκτείνει το μοντέλο βασισμένο στα γεγονότα. Όπως και στον έλεγχο συμμόρφωσης, έτσι και στον επανασχεδιασμό λαμβάνει ως είσοδο και τα δεδομένα και το μοντέλο. Στόχος είναι η αλλαγή του υπάρχοντος μοντέλου. Για παράδειγμα αν κατά τη διάρκεια του ελέγχου συμμόρφωσης διαπιστωθεί πως κάποιο κομμάτι στο μοντέλο δεν ανταποκρίνεται στην πραγματικότητα μπορεί να αλλάξει αυτό το κομμάτι, έτσι ώστε να αντικατοπτρίζει καλύτερα τα πραγματικά γεγονότα. Στην εικόνα που ακολουθεί (Εικόνα 4) φαίνεται πως ενώνοντας τα δεδομένα που παίρνουμε από τα αρχεία καταγραφής με ένα ήδη σχεδιασμένο μοντέλο και τα συμπεράσματα που έχουν βγει από αυτό το μοντέλο δημιουργούμε ένα νέο μοντέλο διαδικασιών.

Εικόνα 4 : Επανασχεδιασμός

Τέλος υπάρχει η επιχειρησιακή υποστήριξη που θα μπορούσε κανείς να πει πως είναι μια εξέλιξη του ελέγχου συμμόρφωσης. Σκοπός της επιχειρησιακής υποστήριξης είναι να επηρεάσει άμεσα τη διαδικασία παρέχοντας προειδοποιήσεις, προβλέψεις ή / και συστάσεις. Ο έλεγχος της συμμόρφωσης μπορεί να γίνει σε πραγματικό χρόνο επιτρέποντας στους ανθρώπους να ενεργούν τη στιγμή που τα πράγματα αποκλίνουν. Με βάση τα δεδομένα μοντέλου και τρεχουσών συμβάντων, μπορεί κανείς να προβλέψει τον υπολειπόμενο χρόνο ροής, τη πιθανότητα τήρησης της προθεσμίας, των σχετικών δαπανών, της πιθανότητας απόρριψης μιας υπόθεσης (όπως η έκδοση ενός δανείου) και ούτω καθεξής. Η διαδικασία δεν βελτιώνεται αλλάζοντας το μοντέλο, αλλά παρέχοντας απευθείας υποστήριξη βάσει δεδομένων με τη μορφή προειδοποιήσεων, προβλέψεων ή / και συστάσεων.

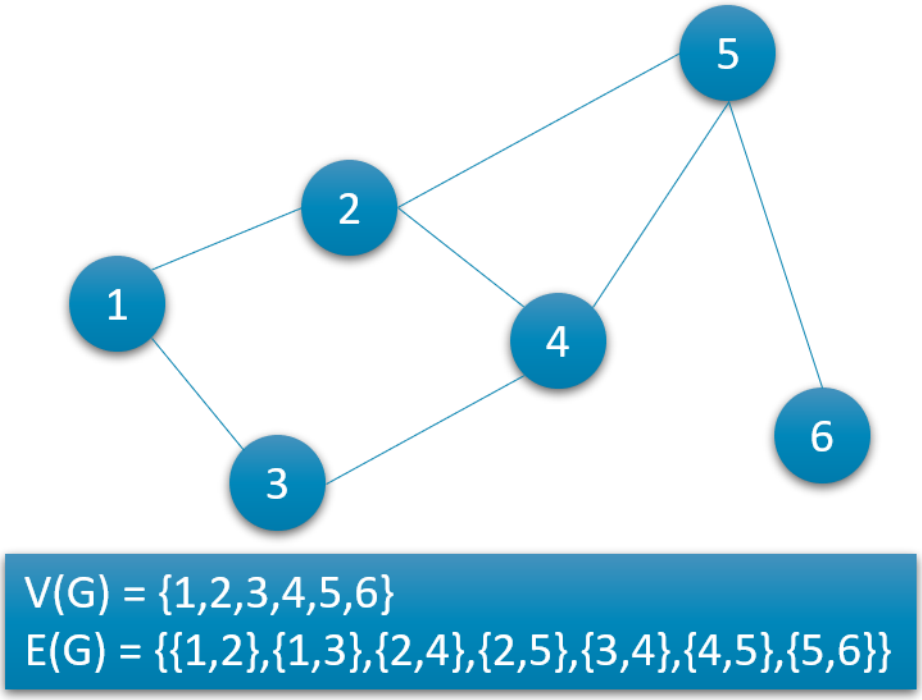
Εικόνα 5 : Επιχειρησιακή υποστήριξη

3 Θεωρία Γραφημάτων – Graph Theory

Στο κεφάλαιο αυτό θα παρουσιάσουμε κάποια βασικά στοιχεία της θεωρίας γραφημάτων, καθώς τα γραφήματα χρησιμοποιούνται για την ανάλυση των σχέσεων και την αναπαράσταση των Bayesian Network. Αρχικά θα ορίσουμε τι είναι ένα γράφημα και πως συμβολίζεται, ενώ στη συνέχεια θα κατηγοριοποιήσουμε τα γραφήματα σε διάφορες κατηγορίες ανάλογα με τις ιδιότητές τους. Τέλος γίνεται αναφορά σε ορισμένες ακολουθίες κόμβων και ακμών.

3.1 Ορισμός και βασικά στοιχεία

Η θεωρία γραφημάτων είναι ο κλάδος των μαθηματικών που ασχολείται με τα γραφήματα. Ένα γράφημα θα μπορούσαμε να το ορίσουμε ως ένα μαθηματικό αντικείμενο που έχει τη δυνατότητα να αναπαρασταθεί εύκολα και απλά με εικόνες. Κάθε γράφημα G αποτελείται από μια δυάδα συνόλων (V,E), όπου το V είναι το σύνολο των κορυφών (vertices) και το E είναι ένα σύνολο ακμών (edges). Η ακμή είναι ένα ζεύγος κορυφών, είτε διατεταγμένων είτε μη-διατεταγμένων ανάλογα με τον τύπο του γραφήματος. Οι ακμές ενός γραφήματος G = (V,E) συμβολίζονται με {x,y} για μη-κατευθυνόμενα γραφήματα και (x,y) για κατευθυνόμενα γραφήματα, όπου x,y κορυφές του γραφήματος G. Εν κατακλείδι ένα γράφημα G είναι ένα μαθηματικό αντικείμενο αυστηρά οριζόμενο από ένα σύνολο κόμβων V(G) και ένα σύνολο ακμών E(G), ενώ για τη αναπαράσταση του γραφήματος με εικόνες αναπαριστούμε με κύκλους τους κόμβους και τις ακμές με ευθείες γραμμές που ενώνουν κόμβους.

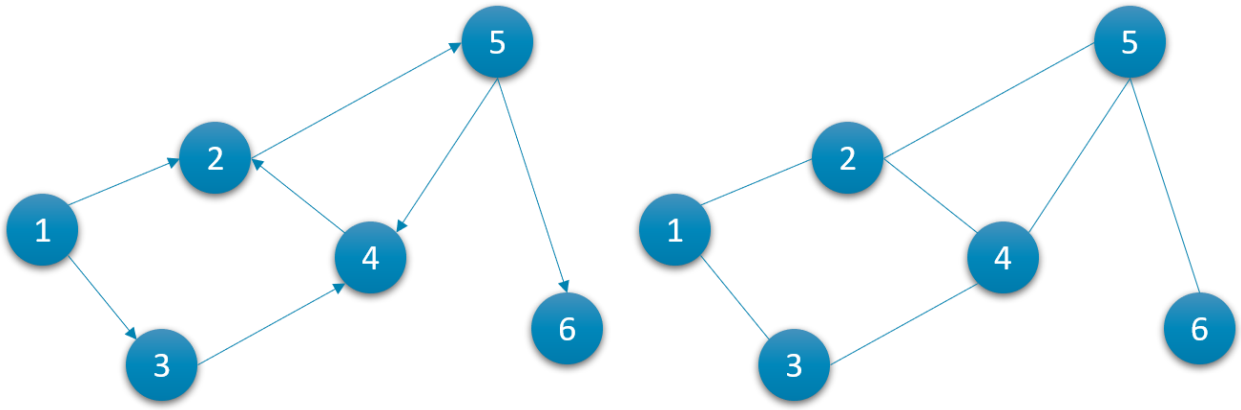


Εικόνα 6 : Μη-κατευθυνόμενο γράφημα με τα σύνολα κορυφών και ακμών του

Όπως γίνεται εύκολα αντιληπτό, από τον ορισμό καθώς και από την παραπάνω εικόνα (Εικόνα 6), ένα γράφημα αποτελεί έναν πολύ εύκολο και φυσικό τρόπο απεικόνισης των σχέσεων μεταξύ αντικειμένων. Για παράδειγμα σε αυτή την εργασία απεικονίζουμε τις διαδικασίες με κόμβους και ενώνουμε τους κόμβους με ακμές για να αναδείξουμε την σχέση μεταξύ τους. Τέτοιες αναπαραστάσεις αντικειμένων και σχέσεων είναι χρήσιμες σε πολλές εφαρμογές για την ανάλυση και την προσέγγιση ενός προβλήματος.

3.2 Κατευθυνόμενα και μη κατευθυνόμενα γραφήματα

Ο τύπος των ακμών ενός γραφήματος καθορίζει αν το γράφημα είναι κατευθυνόμενο ή μη. Όταν το σύνολο E αποτελείται από μη-διατεταγμένα ζεύγη κόμβων ορίζει ένα μη κατευθυνόμενο γράφημα, ενώ στη περίπτωση που το σύνολο E περιέχει διατεταγμένα ζεύγη κόμβων το γράφημα ορίζεται ως κατευθυνόμενο. Τα κατευθυνόμενα γραφήματα ουσιαστικά τονίζουν την κατεύθυνση της σχέσης μεταξύ δυο κόμβων (εικόνα 7).



Εικόνα 7 Κατευθυνόμενο γράφημα (αριστερά), Μη-κατευθυνόμενο γράφημα (δεξιά)

Τάξη (order) γραφήματος καλείται ο αριθμός των κορυφών του γραφήματος και συμβολίζεται με το αγγλικό γράμμα n. Αντίστοιχα μέγεθος ενός γραφήματος ονομάζεται ο αριθμός των ακμών και συμβολίζεται με το αγγλικό γράμμα m. Για το μέγεθος ενός γραφήματος G τάξης n ισχύει πως 0<=m<=n(n-1)/2. Στην περίπτωση που m = 0, τότε το γράφημα ονομάζεται ανεξάρτητο ή ευσταθές, ενώ στην περίπτωση που ισχύει ότι m = n(n-1)/2 τότε το γράφημα ονομάζεται πλήρες.

Ο αριθμός των ακμών που προσπίπτουν σε μια κορυφή καλείται βαθμός (degree) της κορυφής και συμβολίζεται ως $d(\upsilon_i)$. Στην περίπτωση των κατευθυνόμενων γραφημάτων επειδή η κάθε ακμή έχει μια πηγή και έναν προορισμό χωρίζουμε και τον βαθμό τις ακμής σε βαθμός πηγής και βαθμό προορισμού.

Για μη-κατευθυνόμενο γράφημα :

$$\mathbf{d(\upsilon) = |N(\upsilon)|}$$

Για κατευθυνόμενο γράφημα :

$$\mathbf{N^+(v) = \{ u \in V(G) : (v,u) \in E(G)\},\ d^+(v) = |N(v)|}$$

$$\mathbf{N^-(v) = \{ u \in V(G) : (v,u) \in E(G)\},\ d^-(v) = |N(v)|}$$

Για τα γραφήματα της εικόνας 1 έχουμε:

Για το κατευθυνόμενο γράφημα:

$$d^+(1) = 2,\ d^-(1) = 0$$

$$d^+(2) = 1,\ d^-(2) = 2$$

$$d^+(3) = 1,\ d^-(3) = 1$$

$$d^+(4) = 1,\ d^-(4) = 2$$

$$d^+(5) = 2,\ d^-(5) = 1$$

$$d^+(6) = 0,\ d^-(6) = 1$$

Για το μη-κατευθυνόμενο γράφημα:

$$d(1) = 2$$

$$d(2) = 3$$

$$d(3) = 2$$

$$d(4) = 3$$

$$d(5) = 3$$

$$d(6) = 1$$

Δυο βασικά θεωρήματα των γραφημάτων που αξίζουν αναφοράς είναι τα εξής:

1. Το άθροισμα των βαθμών των κορυφών ενός γραφήματος G είναι ίσο με 2m, όπου m είναι ο αριθμός των ακμών του γραφήματος.

Απόδειξη : Αφού κάθε ακμή προσπίπτει σε δυο κορυφές, κατά τον υπολογισμό του αθροίσματος των βαθμών των κορυφών κάθε ακμή προσμετρείται δυο φορές. Επομένως όλες οι ακμές μαζί προσμετρούνται 2m φορές στο άθροισμα των βαθμών του γραφήματος.

2. Σε κάθε γράφημα ο αριθμός των κορυφών με μονό βαθμό είναι ζυγός.

3.3 Βεβαρημένα γραφήματα – Weighted graph

Όταν σε κάθε ακμή e ενός γραφήματος G έχει ανατεθεί μια τιμή w(e), τότε το γράφημα καλείται έμβαρο (weighted) και η τιμή αυτή ονομάζεται βάρος της ακμής e. Ένα έμβαρο γράφημα μπορεί να είναι είτε κατευθυνόμενο είτε μη-κατευθυνόμενο. Ο στόχος των βαρών μπορούν να αντιπροσωπεύσουν πολλά πράγματα, όπως για παράδειγμα αποστάσεις, κόστη ή στην περίπτωση της εξόρυξης διεργασιών μπορεί να δείχνει το πόσες φορές από ένα γεγονός πήγαμε στο άλλο. Συχνά τα έμβαρα γραφήματα αναφέρονται και ως δίκτυα (networks). Το άθροισμα των βαρών όλων των ακμών ενός εμβαρούς γραφήματος G(V,E) ορίζεται ως το βάρος του γραφήματος αυτού.

Ένα γράφημα G'(V',E') ορίζεται ως υπογράφημα ενός άλλου γραφήματος G(V,E) αν το σύνολο των κορυφών του G' είναι υποσύνολο των κορυφών του G και αντίστοιχα το σύνολο ακμών του G' είναι υποσύνολο των ακμών του G. Συμβολίζεται ως G'⊆G όπου V′ ⊆ V, E′ ⊆ E. Ιδιαίτερο ενδιαφέρον παρουσιάζουν δυο τύποι υπογραφημάτων, τα υπογραφήματα που παράγονται από ένα υποσύνολο των ακμών του αρχικού γραφήματος G (spanned subgraph) και τα υπογραφήματα που επάγονται από ένα υποσύνολο κόμβων του αρχικού γραφήματος (induced subgraph).

Έστω S υποσύνολο του συνόλου ακμών E του γραφήματος G. Το παραγόμενο υπογράφημα των ακμών του συνόλου S είναι το γράφημα G' = (V_s,S) για το οποίο ισχύει ότι:

Vs = {x ∈ V | x που είναι κόμβος μιας ακμής του S}

Αντίστοιχα για τα επαγόμενα υπογραφήματα έστω A ένα υποσύνολο κόμβων του γραφήματος, τότε το επαγόμενο υπογράφημα από το σύνολο των κόμβων A είναι το γράφημα G'=(A,E_A), όπου:

E_A = {xy | x ∈ A και y ∈ A}

Ένα υπογράφημα G' καλείται μεγιστοτικό αν δεν υπάρχει άλλο υπογράφημα G'' τέτοιο ώστε G'⊆G''. ΄Ενα επαγόμενο υπογράφημα G′(V′, E′) του G περιέχει κάθε ακμή ανάμεσα στους κόμβους του V′ που υπάρχει στο G. Είναι δηλαδή ένα μεγιστοτικό υπογράφημα του G ως προς V′. Το συμβολίζουμε ως G[V′]. ΄Ενα γεννητορικό υπογράφημα G′(V′, E′) του G(V, E) έχει V = V′ και E′ ⊂ E. ΄Αρα το G′ είναι μεγιστοτικό ως προς το σύνολο E′.

3.4 Ακολουθίες κόμβων και ακμών

Τέλος είναι σημαντικό να αναφερθούμε σε ορισμένες ακολουθίες κόμβων και ακμών. Μια ακολουθία κόμβων $W = (u_1, u_2, u_3, u_4, \ldots, u_n)$ του γραφήματος $G(V,E)$ ονομάζεται περίπατος (walk), όταν $u_{i-1}u_i \in E$ για κάθε $i = 1,2,3,\ldots,n$. Για την απόδοση της έννοιας του περιπάτου συχνά χρησιμοποιούνται οι όροι αλυσίδα (chain) ή ακολουθία ακμών. υ. Το μήκος του περιπάτου W ισούται με το πλήθος των ακμών του, δηλαδή $l(W) = n$.

Εκτός του περιπάτου υπάρχει και το ίχνος (trail), η διαφορά μεταξύ των δυο είναι πως σε ένα ίχνος δεν μπορεί να εμφανιστεί μια ακμή περισσότερες από μια φορές. Με άλλα λόγια μια ακολουθία κόμβων $T = (u_1, u_2, u_3, u_4, \ldots, u_n)$ του γραφήματος G ονομάζεται ίχνος, εάν $u_{i-1}u_i \in E$ για κάθε $i = 1,2,3,\ldots,n$ και δεν υπάρχει ζεύγος διαδοχικών κόμβων του ίχνους T που να εμφανίζεται περισσότερες από μια φορές.

Μια επιπλέον ακολουθία είναι αυτή της διαδρομής ή μονοπατιού (path). Διαδρομή μήκους n ονομάζεται η ακολουθία κόμβων $P = (u_1, u_2, u_3, u_4, \ldots, u_n)$, όταν $u_{i-1}u_i \in E$ για κάθε $i = 1,2,3,\ldots,n$ και δεν υπάρχει κάποιος κόμβος του συνόλου P που να επαναλαμβάνεται, δηλαδή κάθε κόμβος δεν εμφανίζεται περισσότερο από μια φορές.

Η τελευταία ακολουθία που θα αναφέρουμε και άκρως σημαντική για την παρούσα εργασία είναι ο κύκλος (cycle). Κύκλος μήκους n είναι μια ακολουθία κόμβων $C = (u_1, u_2, u_3, u_4, \ldots, u_{n-1}, u_0)$, εάν $u_{i-1}u_i \in E$ για κάθε $i = 1,2,3,\ldots,n-1$ και $u_{n-1}u_0 \in E$, δηλαδή είναι ένας περίπατος όπου μόνο η τερματική κορυφή εμφανίζεται δυο φορές.

4 Πιθανότητες – Probabilities

Η πιθανότητα είναι μια λέξη που χρησιμοποιείται πολύ συχνά στη καθημερινή ζωή, συχνά μιλάμε για την πιθανότητα να βρέξει ή να έχει καλό καιρό. Με άλλα λόγια όταν χρησιμοποιούμε τη λέξη πιθανότητα στην καθημερινή μας ζωή αναφερόμαστε στο βαθμό της σιγουριάς που έχουμε πως κάτι αβέβαιο θα συμβεί. Όμως η έννοια της πιθανότητα έχει πλούσια ιστορία που περιλαμβάνει αρκετές διαφορετικές προσεγγίσεις. Μερικές από αυτές που αξίζει να αναφέρουμε είναι η έννοια της πιθανότητας ως αναλογία (ratio), ως σχετική συχνότητα (relative frequency) και ως βαθμός πεποίθησης (degree of belief). Στο κεφάλαιο αυτό θα περιγράψουμε αυτές τις προσεγγίσεις, θα μιλήσουμε για τα βασικά στοιχεία των πιθανοτήτων και στο τέλος θα εμβαθύνουμε στο κομμάτι των πιθανοτήτων που θα χρειαστούμε για να κατανοήσουμε τα Bayesian Networks.

4.1 Βασικά στοιχεία της θεωρίας πιθανοτήτων

Τα σύνολα (sets) είναι τα θεμέλια πάνω στα οποία στηρίζεται η θεωρία των πιθανοτήτων. Σύμφωνα με τον A.N. Kolmogorov η θεωρία πιθανοτήτων ασχολείται με την πραγματοποίηση πειραμάτων, τα οποία αποτελούνται από σύνολα διακριτών αποτελεσμάτων. Για παράδειγμα η ρίψη ενός ζαριού είναι ένα πείραμα και έχει ένα σύνολο αποτελεσμάτων που είναι οι έξι πλευρές του ζαριού, παρομοίως η ρίψη ενός νομίσματος που μπορεί να φέρει ως αποτέλεσμα κορώνα ή γράμματα. Επιπλέον η επιλογή ενός φοιτητή από το σύνολο φοιτητών ενός μαθήματος και ο προσδιορισμός του αν ο φοιτητής πέρασε το μάθημα ή όχι. Όμως δεν είναι όλα τα πειράματα εύκολα για την εξαγωγή μιας πιθανότητας. Αρχικά σε ένα πείραμα το οποίο δεν είναι καλά ορισμένο δεν μπορείς να ορίσεις μια πιθανότητα. Όταν λέμε καλά ορισμένο εννοούμε για παράδειγμα να μην έχει προσδιοριστεί ένα σύνολο αποτελεσμάτων και επομένως μέχρι να προσδιοριστεί αυτό το σύνολο το πείραμα δεν είναι καλά καθορισμένο. Ένα δεύτερο σημείο που θέλει προσοχή είναι πως μπορεί ένα σύνολο αποτελεσμάτων να είναι άπειρο, όμως στην παρούσα εργασία δεν θα ασχοληθούμε με σύνολα άπειρων αποτελεσμάτων.

Μόλις ένα πείραμα είναι καλά καθορισμένο το σύνολο όλων των αποτελεσμάτων ονομάζεται δειγματικός χώρος (sample space) και συμβολίζεται με το γράμμα “Ω”. Για παράδειγμα σε ένα ζάρι έχουμε ένα σύνολο πιθανών αποτελεσμάτων, τα οποία είναι κάθε πλευρά του ζαριού. Επομένως θέτουμε ως Ω το σύνολο των πιθανών αποτελεσμάτων, δηλαδή $\Omega = \{1,2,3,4,5,6\}$. Στην περίπτωση των πεπερασμένων (finite) δειγματικών χώρων κάθε υποσύνολο του

δειγματικού χώρου ονομάζεται ενδεχόμενο (event) και λέμε πως το Α συνέβη (occurred) όταν το αποτέλεσμα του πειράματος βρίσκεται μέσα στο σύνολο του Α. Ένα υποσύνολο που περιέχει ένα και μόνο ένα στοιχείο αποκαλείται στοιχειώδη στοιχειώδες ενδεχόμενο (elementary event). Στο παραπάνω παράδειγμα με το ζάρι θα μπορούσαμε να πούμε πως το Α = {2,4,6} είναι το ενδεχόμενη η ρίψη του ζαριού να φέρει άρτιο αποτέλεσμα, αν το αποτέλεσμα της ρίψης του ζαριού είναι το 4 λέμε πως το Α συνέβη. Αφού ορίσαμε τον δειγματικό χώρο τώρα μπορούμε να προχωρήσουμε στον ορισμό της συνάρτησης πιθανοτήτων.

Δεδομένου ενός δειγματικού χώρου Ω που περιέχει n διακριτά στοιχεία (elements) , δηλαδή Ω = {e1,e2,...,en}. Μια συνάρτηση που αναθέτει μια πραγματική τιμή P(E) για κάθε ενδεχόμενο E ⊆ Ω καλείται συνάρτηση πιθανοτήτων για το σύνολο των υποσυνόλων του Ω αν αυτό ικανοποιεί τις ακόλουθες συνθήκες :

1. $0 \leq P(\{ei\}) \leq 1$, για κάθε i όπου $1 \leq i \leq n$.

2. $P(\{e1\}) + P(\{e2\}) + ... + P(\{en\})=1$.

3. Για κάθε ενδεχόμενο E = {ei1 , ei2 ,...eik }, το οποίο δεν είναι στοιχειώδες ενδεχόμενο, ισχύει : $P(E) = P(\{ei1\}) + P(\{ei2\}) + ... + P(\{eik\})$.

Το ζευγάρι (Ω,P) καλείται πιθανοτικός χώρος.

Έστω ένα πείραμα ρίψης ενός νομίσματος. Ο δειγματικός χώρος Ω αποτελείται από τα δύο πιθανά ενδεχόμενα, δηλαδή κορώνα και γράμματα. Θεωρώντας πως όλα τα ενδεχόμενα είναι ισοπίθανα μπορούμε να αναθέσουμε την πιθανότητα $P(\{e\}) = \frac{1}{2}$ για κάθε e ∈ Ω. Και άρα θέτοντας ως ενδεχόμενο Α το νόμισμα να φέρει κορώνα και ενδεχόμενο Β το νόμισμα να φέρει γράμματα έχουμε $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$ και $P(A, B) = \frac{2}{2}$.

Δύο στοιχειώδη ενδεχόμενα λέμε πως είναι ισοπίθανα όταν δεν έχουμε λόγο να πιστεύουμε πως το ένα είναι πιο πιθανό να συμβεί από το άλλο. Σύμφωνα με αυτή τη λογική για ένα πείραμα n στοιχείων η πιθανότητα του καθενός να συμβεί είναι ίση με το $\frac{1}{n}$. Αυτή η προσέγγιση για την ανάθεση πιθανοτήτων καλείται αναλογία (ratio) και συχνά χρησιμοποιείται στην ανάθεση πιθανοτήτων για παιχνίδια τύχης όπως η ρίψη ζαριού για παράδειγμα.

Έστω πως έχουμε όμως ένα "πειραγμένο" ζάρι του οποίου κάποιες πλευρές είναι πιο πιθανό να έρθουν από κάποιες άλλες. Λόγω αυτής της ιδιαιτερότητας καταλαβαίνουμε πως τα ενδεχόμενα του πειράματος δεν είναι ισοπίθανα, επομένως δεν μπορούμε να υπολογίσουμε σύμφωνα με την έννοια της αναλογίας. Για να αντιμετωπιστούν αυτές οι περιπτώσεις το 1919 ο Richard von Mises ανέπτυξε την προσέγγιση σχετικής συχνότητας (relative frequency) σύμφωνα με την οποία αν ένα πείραμα επαναληφθεί πολλές φορές, η πιθανότητα κάθε ενδεχομένου είναι το όριο όταν ο αριθμός των επαναλήψεων του πειράματος τείνει στο άπειρο του αριθμού των εμφανίσεων του ενδεχομένου προς τον αριθμό των επαναλήψεων. Συνεχίζοντας το παράδειγμα του πειραγμένου ζαριού και αν ρίξουμε το ζάρι m φορές τότε η πιθανότητα το ζάρι να φέρει τέσσερα είναι

$$P(\{4\}) = \lim_{m\to\infty} \frac{αριθμός\ εμφάνισης\ του\ 4}{m}$$

Άρα, αν ρίξουμε το ζάρι 10 φορές και 5 φορές το ζάρι φέρει 4 η πιθανότητα μια ρίψη του ζαριού να φέρει 4 είναι 0.5. Σύμφωνα με αυτή τη προσέγγιση, η πιθανότητα που προκύπτει δεν είναι ιδιότητα μια δοκιμής ενός πειράματος αλλά ιδιότητα ενός συνόλου δοκιμών.

Τι γίνεται όμως όταν θέλουμε να υπολογίσουμε μια πιθανότητα για κάτι δεν έχει να κάνει με αναλογίες ή με επαναλήψεις πειραμάτων; Για παράδειγμα όταν εκτιμούμε πως σήμερα το απόγευμα θα βρέξει. Αυτή η εκτίμηση δεν είναι προϊόν πειραμάτων ή αναλογιών, είναι απλά ένας βαθμός πεποίθησης (degree of belief), πιστεύουμε δηλαδή πως υπάρχει η πιθανότητα να βρέξει ή να έχει λιακάδα. Στην παρούσα εργασία καθώς και στο τομέα των Μπεϋζιανών δικτύων χρησιμοποιείται αυτή η έννοια της πιθανότητα, δηλαδή του βαθμού πεποίθησης.

Για ένα πιθανοτικό χώρο (Ω,P) ισχύουν τα εξής (A.N. Kolmogorov in 1933):

1. $P(\Omega) = 1$.
2. $0 \leq P(E) \leq 1$, για κάθε $E \subseteq \Omega$.
3. Για E και $F \subseteq \Omega$ τέτοιο ώστε $E \cap F = \emptyset$, έχουμε πως $P(E \cup F) = P(E) + P(F)$.

4.2 Δεσμευμένη πιθανότητα και ανεξαρτησία

Μια πολύ σημαντική έννοια στο κλάδο των πιθανοτήτων που είναι και βασικό στοιχείο των Μπευζιανών δικτύων είναι η δεσμευμένη πιθανότητα. Με απλά λόγια όταν λέμε δεσμευμένη πιθανότητα εννοούμε τη πιθανότητα του να συμβεί ένα ενδεχόμενο Α ξέροντας πως έχει συμβεί ένα ενδεχόμενο Β. Για παράδειγμα ποια η πιθανότητα του δρόμου να είναι βρεγμένος ξέροντας πως πριν έβρεχε.

Αρχικά η έννοια της δεσμευμένης πιθανότητας ξεκίνησε κατά την εκτίμηση πιθανοτήτων ως αναλογίες. Στην περίπτωση των αναλογιών η δεσμευμένη πιθανότητα του Α δεδομένου του ενδεχομένου Β, δηλαδή P(A|B), είναι το κλάσμα των στοιχείων του ενδεχομένου Β που ανήκουν ταυτόχρονα και στο ενδεχόμενο Α. Έστω Α και Β δύο ενδεχόμενα τέτοια ώστε P(B) ≠ 0. Θέτουμε ως δεσμευμένη πιθανότητα του Α δεδομένου του ενδεχομένου Β ως P(A|B) και ορίζεται ως εξής :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(Neapolitan, R. E. (2004). *Learning bayesian networks* (Vol. 38). Upper Saddle River, NJ: Pearson Prentice Hall.)

Πολύ συχνά χρησιμοποιούμε τον όρο ανεξάρτητα ενδεχόμενα, το οποίο δηλώνει πως δύο ενδεχόμενα Α και Β είναι ανεξάρτητα μεταξύ τους όταν το αποτέλεσμα του ενός δεν επηρεάζει το άλλο. Αν σε μια τράπουλα τραβήξουμε ένα φύλλο και μας πουν πως το φύλλο που τραβήξαμε είναι σπαθί δεν επηρεάζει το αν το φύλλο είναι Βαλές ή Ρίγας. Ακολουθεί ο ορισμός των ανεξάρτητων ενδεχομένων.

Έστω τα ενδεχόμενα Α και Β, λέμε πως τα δύο αυτά ενδεχόμενα είναι ανεξάρτητα όταν :

1. P(A|B) = P(A) και P(A) ≠ 0, P(B) ≠ 0.
2. P(A)=0 ή P(B)=0.

Επιπλέον δύο ενδεχόμενα Α και Β είναι υπό συνθήκη ανεξάρτητα όταν δεδομένου ενός ενδεχομένου Γ, όπου ισχύει πως P(Γ) ≠ 0 και τουλάχιστον ένα από τα παρακάτω :

1. P(A|B ∩ Γ) = P(A|B) και P(A|Γ) ≠ 0, P(B|Γ) ≠ 0.
2. P(A|Γ) = 0 ή P(B|Γ) = 0.

Ένας πολύ χρήσιμος κανόνας των δεσμευμένων πιθανοτήτων είναι ο εξής, έστω n ενδεχόμενα ($E_1,E_2,E_3,...,E_n$) τέτοια ώστε $E_i \cap E_j = Ø$ για κάθε $i \neq j$ και $E_1 \cup E_2 \cup E_3 \cup ... \cup E_n = Ω$. Αυτά τα ενδεχόμενα ονομάζονται αμοιβαία αποκλειόμενα. Στην περίπτωση αυτή το θεώρημα ολικής πιθανότητας (law of total probability) αναφέρει πως για κάθε ενδεχόμενο A έχουμε :

$$P(A) = \sum_{i=1}^{n} P(A \cap Ei$$

Αν $P(E_i) \neq 0$, τότε $P(A \cap E_i) = P(A|E_i)P(E_i)$. Επομένως όταν $P(E_i) \neq 0$ για κάθε i το θεώρημα παίρνει την εξής μορφή:

$$P(A) = \sum_{i=1}^{n} P(A|Ei)P(Ei)$$

4.3 Θεώρημα του Bayes

Αφού ορίσαμε τη δεσμευμένη πιθανότητα είναι πλέον εύκολο να προχωρήσουμε σε θεωρήματα που πηγάζουν από αυτόν τον απλό ορισμό. Αρχικά μπορούμε να βγάλουμε έναν πιο απλό τύπο για τον υπολογισμό της τομής δυο ενδεχομένων, πολλαπλασιάζοντας και τα δυο μέρη της εξίσωσης της δεσμευμένης πιθανότητας με το παρονομαστή. Δηλαδή έχουμε $P(A \cap B)=P(B \mid A)\ P(A)$. Γενικεύοντας το παραπάνω για την τομή «n» ενδεχομένων προκύπτει :

$$P(A_1,A_2,...,A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1,A_2)...P(A_n|A_1,...,A_{n-1})$$

Πλέον είμαστε έτοιμοι να ορίσουμε τον κανόνα του Bayes, ο οποίος είναι πολύ σημαντικός σε διάφορες εφαρμογές πιθανοτήτων και στατιστικής.

Έστω δύο ενδεχόμενα A και B τέτοια ώστε $P(A) \neq 0$ και $P(B) \neq 0$, τότε έχουμε:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Επιπλέον, αν έχουμε n αμοιβαία αποκλειόμενα ενδεχόμενα $A_1,A_2,...,A_n$, τέτοια ώστε $P(A_i) \neq 0$ για κάθε i όπου $1 \leq i \leq n$,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)\mathrm{P}(\mathrm{A}_1) + P(B|A_2)\mathrm{P}(\mathrm{A}_2) + \ldots + P(B|A_n)\mathrm{P}(\mathrm{A}_n)}$$

Και οι δυο παραπάνω εξισώσεις αναφέρονται ως το θεώρημα Bayes και αναπτύχθηκαν από τον Thomas Bayes. Με τη χρήση της πρώτης εξίσωσης μπορούμε να υπολογίσουμε το P(A|B), αν γνωρίζουμε το P(B|A), το P(A) και το P(B), ενώ με την δεύτερη εξίσωση μας δίνεται η δυνατότητα να υπολογίσουμε το $\mathrm{P}(A_i|B)$ αν γνωρίζουμε το $\mathrm{P}(\mathrm{B}|\mathrm{A_i})$, το $\mathrm{P}(\mathrm{A_i})$ για $1 \leq i \leq n$. Ο υπολογισμός δεσμευμένων πιθανοτήτων χρησιμοποιώντας κάποια από τις παραπάνω εξισώσεις.

4.4 Τυχαίες μεταβλητές και κατανομή από κοινού πιθανοτήτων

Κατά την εκτέλεση ενός πειράματος συχνά ο στόχος είναι η εύρεση κάποιας συνάρτησης του αποτελέσματος και όχι το ίδιο το αποτέλεσμα. Για παράδειγμα, σε ένα πείραμα ρίψης δυο ζαριών μπορεί να μας ενδιαφέρει να δούμε ποιο είναι το άθροισμα των τιμών των δυο ζαριών και όχι την ξεχωριστή τιμή που έφερε κάθε ζάρι. Δηλαδή αν θέλουμε να δούμε αν το άθροισμα είναι 5 δεν μας απασχολεί αν αυτό το άθροισμα προέρχεται από το συνδυασμό {3,2} ή {4,1}. Επίσης στο στρίψιμο ενός νομίσματος είναι πιθανό να ψάχνουμε τον συνολικό αριθμό των φορών που το νόμισμα έφερε κορώνα και όχι την ακολουθία αποτελεσμάτων των μεμονωμένων ρίψεων. Από μαθηματικής άποψης, αυτές οι ενδιαφέρουσες ποσότητες είναι συναρτήσεις με πεδίο ορισμού τον δειγματικό χώρο και τιμές πραγματικούς αριθμούς, και ονομάζονται τυχαίες μεταβλητές [7]. Επειδή η τιμή μιας τυχαίας μεταβλητής προσδιορίζεται από το αποτέλεσμα του πειράματος, μπορούμε να αντιστοιχίσουμε πιθανότητες στις δυνατές τιμές της τυχαίας μεταβλητής.

Ορίζουμε ως τυχαία μεταβλητή X μια συνάρτηση του δειγματικού χώρου Ω, δεδομένου ενός πιθανοτικού χώρου (Ω,P). Αυτή η τυχαία μεταβλητή αναθέτει μια μοναδική τιμή για κάθε στοιχείο (ενδεχόμενο) του δειγματικού χώρου. Το σύνολο των τιμών που ανέθεσε μια τυχαία μεταβλητή X, ονομάζεται χώρος του X. Όταν ο χώρος είναι πεπερασμένος ή μετρήσιμος τότε η τυχαία μεταβλητή θα λέμε πως είναι διακριτή. Στην συνέχεια παρουσιάζεται ένα παράδειγμα χρήσης τυχαίων μεταβλητών.

Ας πάρουμε για παράδειγμα τη ρίψη δύο κανονικών ζαριών, δηλαδή κάθε ενδεχόμενο-αποτέλεσμα έχει πιθανότητα να συμβεί P(ε) = 1/36. Ο δειγματικός χώρος Ω είναι το παρακάτω σύνολο ταξινομημένων ζευγαριών για τα αποτελέσματα των δύο ζαριών :

$$\Omega = \{(1, 1),(1, 2),(1, 3),(1, 4),(1, 5),(1, 6),(2, 1),(2, 2),...(6, 5),(6, 6)\}.$$

Θέτουμε ως τυχαία μεταβλητή X το άθροισμα των αποτελεσμάτων των δύο ζαριών μετά από μια ρίψη και δημιουργούμε τον παρακάτω πίνακα:

Ενδεχόμενο e	X(e)
(1,1)	2
(1,2)	3
…	…
(6,6)	12

Ο χώρος της μεταβλητής X είναι το σύνολο {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}.

Για κάθε τυχαία μεταβλητή X, χρησιμοποιούμε X = x για να δηλώσουμε το σύνολο των στοιχείων e ∈ Ω τα οποία η μεταβλητή X αντιστοιχεί σε μια τιμή x.Το X = x αντιπροσωπεύει το ενδεχόμενο {e για το οποίο X(e) = x}. Σημαντικό είναι να τονιστεί πως με το κεφαλαίο X αντιπροσωπεύουμε τη συνάρτηση ενώ με το μικρό x δηλώνουμε οποιοδήποτε ενδεχόμενο που ανήκει στον χώρο του X.

Έστω οι τυχαίες μεταβλητές X,Y, ορισμένες σε έναν πιθανοτικό χώρο, η κατανομή από κοινού πιθανοτήτων για τις X,Y είναι η πιθανοτική κατανομή που αναθέτει πιθανότητες στις τυχαίες μεταβλητές για κάθε πιθανό σύνολο τιμών που αυτές μπορούν να πάρουν.

Πιο πάνω ορίσαμε τα ανεξάρτητα ενδεχόμενα, τώρα ήρθε η ώρα να δούμε την ανεξαρτησία συνόλων τυχαίων μεταβλητών. Έστω ένας πιθανοτικός χώρος (Ω,P) και δυο σύνολα A και B που περιέχουν τυχαίες μεταβλητές ορισμένες στο Ω. Τα σύνολα A και B λέμε πως είναι ανεξάρτητα αν για κάθε τιμή των μεταβλητών που υπάρχουν στα σύνολα a και b, τα ενδεχόμενα A = a και B = b είναι ανεξάρτητα. Δηλαδή P(a) = 0 ή P(b) = 0 ή P(a|b) = P(a). Σε αυτή τη περίπτωση γράφουμε $I_p(A,B)$, όπου το I_p σημαίνει ανεξάρτητο στο P.

Για παράδειγμα έστω ο δειγματικός χώρος Ω όπου είναι το σύνολο όλων των καρτών μιας τράπουλας και έστω P = 1/52 η πιθανότητα εμφάνισης κάθε κάρτας. Ορίζουμε τις εξής τυχαίες μεταβλητές :

Μεταβλητή	Τιμή	Αποτελέσματα της τιμής
R	r1	Όλες οι φιγούρες
	r2	Όλες οι κάρτες που δεν είναι φιγούρες
T	t1	Τα δεκάρια και οι βαλέδες
	t2	Οι κάρτες που δεν είναι δεκάρια ή βαλέδες
S	s1	Τα μπαστούνια
	s2	Όχι τα μπαστούνια

Τα σύνολα {R,T} και {S} είναι ανεξάρτητα. Δηλαδή $I_p(\{R,T\},\{S\})$. Για να το αποδείξουμε θα πρέπει για όλες τις τιμές των r,t και s να ισχύει πως P(r,t|s) = P(r,t).

s	r	t	P(r,t|s)	P(r,t)
s1	r1	t1	1/13	1/13
s1	r1	t2	2/13	2/13
s1	r2	t1	1/13	1/13
s1	r2	t2	9/13	9/13
s2	r1	t1	1/13	1/13
s2	r1	t2	2/13	2/13
s2	r2	t1	1/13	1/13
s2	r2	t2	9/13	9/13

Επομένως αποδείξαμε πως τα σύνολα {R,T} και {S} είναι ανεξάρτητα.

Αντίστοιχα στην περίπτωση που έχουμε τρία σύνολα A,B και C που περιέχουν τυχαίες μεταβλητές ορισμένες στο Ω, τότε τα σύνολα A και B λέμε πως είναι υπό όρους ανεξάρτητα δεδομένου ενός συνόλου C αν όλες οι τιμές των μεταβλητών που ανήκουν στα σύνολα a,b και c

όποτε $P(c) \neq 0$ τα ενδεχόμενα $A = a$ και $B = b$ είναι υπό όρους ανεξάρτητα δεδομένου του ενδεχομένου $C = c$. Δηλαδή $P(a|c) = 0$ ή $P(b|c) = 0$ ή $P(a|b,c) = P(a|c)$. Στην περίπτωση αυτή γράφουμε $I_p(A,B|C)$.

5 Γραφικά Μοντέλα – Graphical Models

5.1 Γενικά στοιχεία

Ένα γραφικό μοντέλο ή ένα πιθανοτικό γραφικό μοντέλο (PGM) είναι ένα μοντέλο πιθανοτήτων όπου ένα γράφημα εκφράζει τη δομή υπό όρους εξαρτήσεων μεταξύ τυχαίων μεταβλητών ενός προβλήματος. Χρησιμοποιείται κυρίων στην θεωρία των πιθανοτήτων, στη στατιστική – ιδιαίτερα στην Μπεϋζιανή στατιστική – και στη μηχανική μάθηση.

Ένα γράφημα αποτελείται από ένα σύνολο κόμβων-κορυφών , οι οποίοι στα γραφικά μοντέλα αναπαριστούν μεταβλητές, και ένα σύνολο ακμών. Κάθε ακμή συνδέει δυο κόμβους, στην περίπτωση των κατευθυνόμενων γραφημάτων η ακμή απεικονίζει και μια κατεύθυνση. Η ακμή υποδηλώνει μια σχέση ανάμεσα στις δυο κορυφές

Δυο μεταβλητές που είναι υπό όρους ανεξάρτητες δεν έχουν άμεσο αντίκτυπο η μια στις τιμές της άλλης. Για παράδειγμα, αν η μεταβλητή A είναι υπό όρους ανεξάρτητη από την μεταβλητή B δεδομένης μιας μεταβλητής C ισχύει πως $P(A|C,B) = P(A|C)$ και μπορούμε να το συμβολίσουμε ως $(A \perp\!\!\!\perp B|C)$. Η έννοια της ανεξαρτησίας υπό όρους είναι απαραίτητη για τις θεωρίες στατιστικών συμπερασμάτων που βασίζονται σε γραφήματα, καθώς καθιερώνει μια μαθηματική σχέση μεταξύ μιας συλλογής υπό όρους καταστάσεων και ενός γραφοειδούς.

5.2 Δίκτυα Petri – Petri nets

Τον τελευταίο αιώνα υπάρχει πληθώρα νέων τεχνικών μοντελοποίησης διαδικασιών. Τα δίκτυα Petri διαδραματίζουν ακόμη πιο σημαντικό ρόλο στο BPM καθώς είναι γραφικά και μπορούν να μοντελοποιήσουν τον «συγχρονισμό» (concurrency). Στην πραγματικότητα, οι περισσότερες από τις σύγχρονες σημειώσεις και συστήματα BPM χρησιμοποιούν σημασιολογία βασισμένη σε tokens που έχουν υιοθετηθεί από τα δίκτυα Petri. Το 1962 ο Carl Adam Petri (1926-2010) προτείνει τα Petri nets και βάζει επίσημα στο προσκήνιο για πρώτη φορά τον συγχρονισμό, διότι στις επιχειρησιακές διαδικασίες πολλά πράγματα γίνονται παράλληλα. Επομένως τα συστήματα διαχείρισης επιχειρησιακών διαδικασιών θα πρέπει να υποστηρίζουν τον συγχρονισμό.

Ένα Petri net είναι ένα κατευθυνόμενο διμερές γράφημα (directed bipartite graph) όπου οι κόμβοι κατηγοριοποιούνται σε δύο τύπους, τις καταστάσεις (states) και τις μεταβάσεις (transitions) [4]. Το γράφημα είναι διμερές καθώς οι κόμβοι συνδέονται μεταξύ τους με κατευθυνόμενες ακμές με την προϋπόθεση πως δεν υπάρχει σύνδεση μεταξύ δυο κόμβων του

ίδιου τύπου. Στο γράφημα οι καταστάσεις συμβολίζονται με κύκλους και οι μεταβάσεις συμβολίζονται με ορθογώνια παραλληλόγραμμα. Ακολουθεί ο ορισμός ενός Petri net :

Ένα Petri Net είναι μια πλειάδα τριών ορισμάτων (S,T,F), όπου :

- Το S είναι ένα πεπερασμένο σύνολο (finite set) καταστάσεων.
- Το T είναι ένα πεπερασμένο σύνολο μεταβάσεων.
- Τα S και T θα πρέπει να ανεξάρτητα (disjoint), δηλαδή ένας κόμβος δεν μπορεί να ανήκει και στους δυο τύπους. Θα πρέπει να ισχύει πως $(S \cap T) = \emptyset$.
- $F \subseteq (P \times T) \cup (P \times T)$ είναι ένα σύνολο ακμών, το οποίο ονομάζεται «σχέση ροής» (flow relation).

Επιπλέον από τον ορισμό μπορούμε να εξάγουμε τα εξής:

- Μια κατάσταση s ονομάζεται θέση εισόδου μιας μετάβασης t αν και μόνο αν υπάρχει μια κατευθυνόμενη ακμή από την s προς την t.
- Αντίστοιχα μια κατάσταση s θεωρείται θέση εξόδου μιας μετάβασης t αν και μόνο αν υπάρχει μια κατευθυνόμενη ακμή από την t προς την s.

5.3 Δίκτυα ροής – Workflow Nets

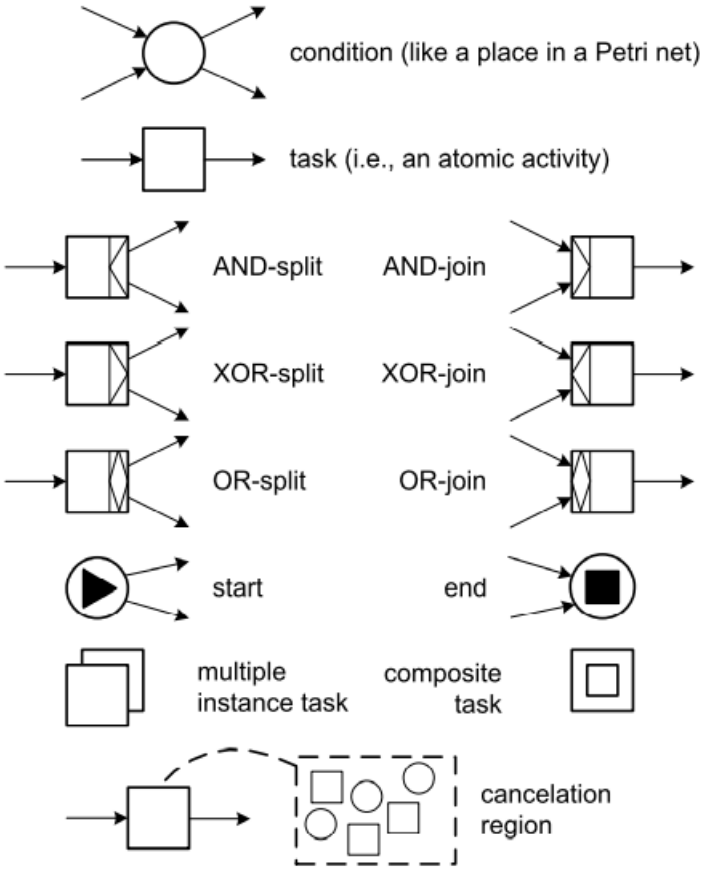
Συχνά κατά τη μοντελοποίηση επιχειρησιακών διαδικασιών χρησιμοποιείται μια υπό-κατηγορία των δικτύων petri που ονομάζονται δίκτυα ροής (workflow nets) και συνήθως τα βλέπουμε γραμμένα ως WF-nets. Κάθε WF-net είναι ένα petri net το οποίο όμως έχει έναν συγκεκριμένο κόμβο πηγή από όπου ξεκινάει η διαδικασία και έναν κόμβο τέλους. Επομένως κάθε κόμβος ανήκει σε ένα μονοπάτι από την πηγή στο τέλος.

Ο λόγος που κάνει ιδιαίτερα σημαντικά τα δίκτυα WF είναι πως παρομοιάζουν τέλεια το κύκλο ζωής μιας διαδικασίας που περιγράψαμε στην διαχείριση επιχειρηματικών διαδικασιών. Για παράδειγμα διαδικασίας μπορούμε να σκεφτούμε την έκδοση δανείου, το κλείσιμο ενός ραντεβού στο γιατρό και την διαδικασία πρόσληψης ενός νέου υπαλλήλου. Σε όλα αυτά τα παραδείγματα έχουμε μια συγκεκριμένη αρχή και ένα συγκεκριμένο τέλος, ανάμεσα σε αυτά τα σημεία υπάρχουν οι δραστηριότητες που αποτελούν την διαδικασία. Ένα μοντέλο μπορεί να εκτελείται πολλές φορές, ως παράδειγμα μπορούμε να σκεφτούμε πως μια διαδικασία για την έκδοση δανείου μπορεί να εκτελείται εκατοντάδες φορές την μέρα από μια τράπεζα. Κάθε τέτοια εκτέλεση θεωρείται ουσιαστικά πως είναι αντίγραφο του ίδιου δικτύου WF.

Τα δίκτυα WF είναι επίσης ένας τρόπος αναπαράστασης που ταιριάζει στην εξόρυξη διαδικασιών. Είναι εμφανής η ομοιότητα μεταξύ ενός δικτύου WF και ενός ίχνους (trace) που βρίσκει κανείς στα αρχεία καταγραφής γεγονότων.

5.4 YAWL

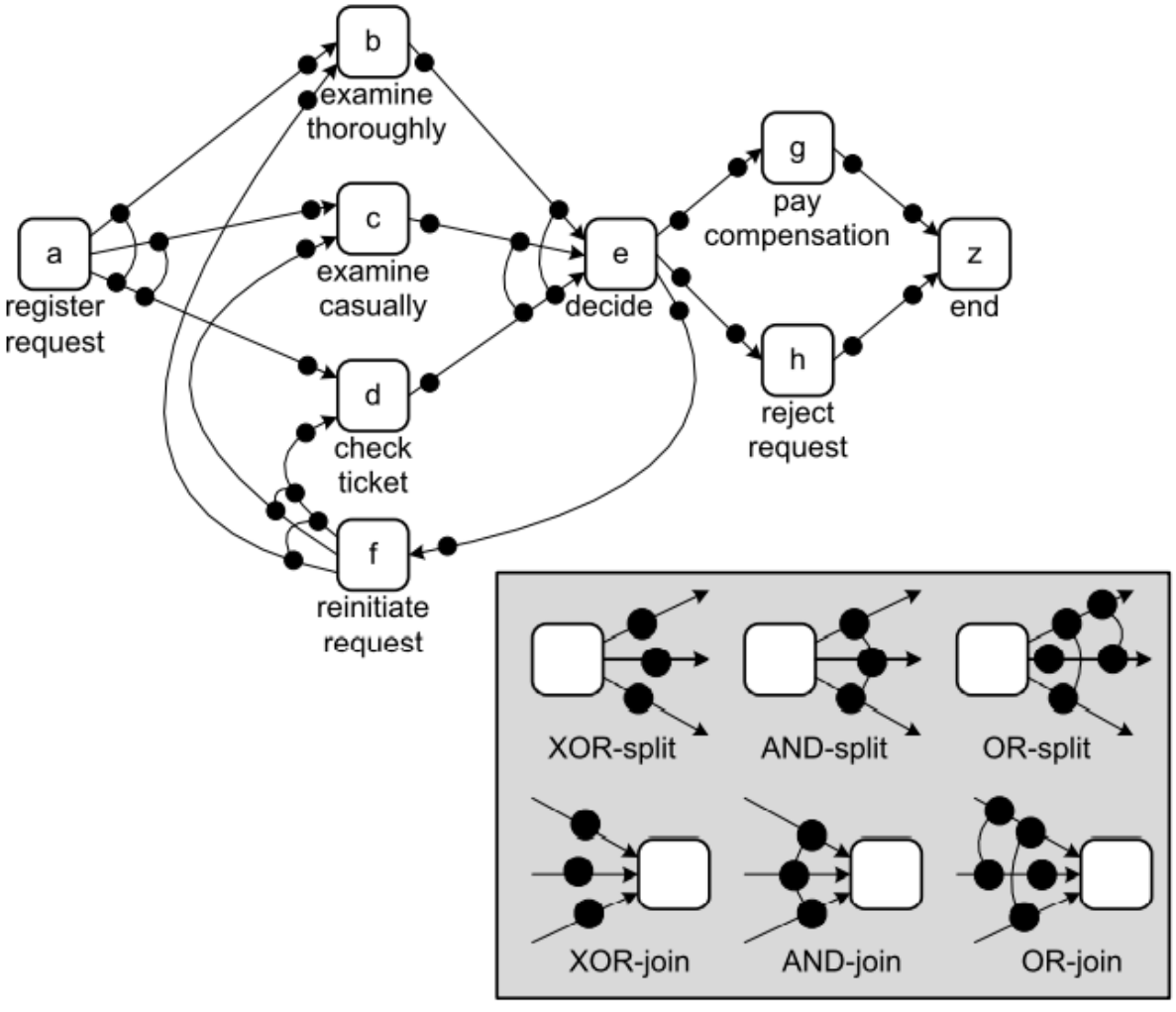
Το YAWL είναι ακρώνυμο για τη φράση «Yet Another Workflow Language» και είναι μια γλώσσα μοντελοποίησης ροής εργασίας. Ο στόχος της συγκεκριμένης γλώσσας είναι να προσφέρει άμεση υποστήριξη πολλών μοτίβων χωρίς όμως η γλώσσα να γίνεται πολύπλοκη. Όπως και στα WF δίκτυα έτσι και το YAWL έχει πάντα εξατομικευμένη αρχική και τελική συνθήκη. Οι συνθήκες του YAWL αντιστοιχούν στα places του petri net. Το YAWL προσφέρει μεγάλη γκάμα σημειογραφιών για πράξεις μεταξύ κόμβων, όπως AND,OR, XOR κοκ. Στην παρακάτω εικόνα (Εικόνα 8) παρουσιάζονται η σημειογραφία που χρησιμοποιείται στη γλώσσα YAWL.



Εικόνα 8 : Σημειογραφία τελεστών

5.5 Causal Nets

Τα Causal Nets είναι μια γραφική αναπαράσταση προσαρμοσμένη πάνω στη εξόρυξη διεργασιών. Ένα causal net είναι ένα γράφημα που περιέχει κόμβους για να αναπαραστήσει τις δραστηριότητες (activities) και ακμές για να αναπαραστήσει σχέσεις μεταξύ των δραστηριοτήτων. Κάθε δραστηριότητα έχει ένα σύνολο πιθανών εισόδων και ένα σύνολο πιθανών εξόδων. Στην εικόνα 9 παρουσιάζεται ένα Causal Net με τις σημειογραφίες όλων των πιθανών τελεστών.



Εικόνα 9 : Causal Nets και σημειογραφία

Ένα Causal net (C-net) ορίζεται ως η πλειάδα $C = (A,a_i,a_o,D,I,O)$ όπου:

- A είναι ένα πεπερασμένο σύνολο δραστηριοτήτων,
- a_i ανήκει στο A και είναι η δραστηριότητα εκκίνησης,
- a_o ανήκει στο A και είναι η δραστηριότητα τερματισμού,
- D είναι η σχέση εξάρτησης,
- I είναι ένα σύνολο των πιθανών εισόδων για κάθε δραστηριότητα,
- O είναι ένα σύνολο των πιθανών εξόδων για κάθε δραστηριότητα.

6 Μπεϋζιανά δίκτυα – Bayesian Networks

Στα προηγούμενα κεφάλαια αναφερθήκαμε σε βασικά κομμάτια των πιθανοτήτων, των γραφημάτων και των γραφικών μοντέλων έτσι ώστε να έχουμε τις βασικές γνώσεις για να μπορέσουμε να προχωρήσουμε πλέον στα Μπεϋζιανά δίκτυα (Bayesian Networks). Στο κεφάλαιο αυτό λοιπόν θα μιλήσουμε για τα Μπεϋζιανά δίκτυα ξεκινώντας από κάποιους εναλλακτικούς ορισμούς των τυχαίων μεταβλητών και των κοινών πιθανοτήτων, έτσι ώστε να είναι πιο εύκολη η εφαρμογή τους στον Μπεϋζιανό συμπερασμό (Bayesian inference). Στη συνέχεια γίνεται λόγος για τις αλυσίδες Μαρκόφ (Markov chains) και ορίζουμε τι είναι ένα δίκτυο Μπέυζ. Ενώ στο τέλος γίνεται αναφορά στην εκμάθηση δομής (structure learning) που είναι ένα σημείο κλειδί για το πρακτικό μέρος της εργασίας.

6.1 Εισαγωγή

Πολλές φορές χρειάζεται να υπολογίσουμε την πιθανότητα ενός αβέβαιου γεγονότος βασιζόμενοι σε κάποια στοιχεία που παρατηρήσαμε. Για παράδειγμα, αν θέλαμε να ξέρουμε ποια είναι η πιθανότητα μιας συγκεκριμένης ασθένειας όταν παρατηρούμε τα συμπτώματα ενός ασθενούς. Τέτοια προβλήματα είναι συχνά πολύπλοκα με πολλές αλληλένδετες μεταβλητές. Μπορεί να οφείλονται σε πολλά συμπτώματα και ακόμη πιο πολλές αιτίες. Συνήθως στην πράξη είναι πιο εύκολο να υπολογιστεί μόνο η αντίστροφη υπό όρους πιθανότητα (reversed conditional probability), όπως για παράδειγμα είναι πολύ πιο εύκολο να παρατηρήσει κανείς τα συμπτώματα ενός ασθενούς ξέροντας πως έχει μια συγκεκριμένη ασθένεια. Σε αυτές τις περιπτώσεις η Μπεϋζιανή σκέψη είναι η κατάλληλη προσέγγιση.

Ένα Μπεϋζιανό δίκτυο (Bayesian Network) αντιπροσωπεύει τις πιθανοτικές σχέσεις αιτιών μεταξύ ενός συνόλου τυχαίων μεταβλητών, τις υπό όρους εξαρτήσεις τους και επιπλέον παρέχει μια συμπαγή αναπαράσταση της κοινής κατανομής πιθανοτήτων. Τα Μπεϋζιανά δίκτυα, επίσης γνωστά ως δίκτυα πεποιθήσεων (belief networks), ανήκουν στην οικογένεια των πιθανοτικών γραφικών μοντέλων. Κάθε Bayesian Network αποτελείται από δυο μέρη, ένα άκυκλο κατευθυνόμενο γράφημα και ένα σύνολο κατανομών υπό όρους πιθανοτήτων. Όταν υπάρχει μια σχέση ανάμεσα σε τυχαίες μεταβλητές του γραφήματος, οι αντίστοιχοι κόμβοι συνδέονται μεταξύ τους με μια κατευθυνόμενη ακμή. Επίσης η κατευθυνόμενη ακμή από έναν κόμβο Α σε έναν κόμβο Β υποδηλώνει ότι η τυχαία μεταβλητή Α προκαλεί την τυχαία μεταβλητή Β. Σε ένα

Μπεϋζιανό δίκτυο οι κύκλοι δεν επιτρέπονται, διότι οι κατευθυνόμενες ακμές αντιπροσωπεύουν μια στατική πιθανότητα εξάρτησης αιτίας.

6.2 Τυχαίες μεταβλητές και από κοινού πιθανότητες για τα Μπεϋζιανό συμπερασμό

Αν και ο ορισμός που αναφέραμε στο κεφάλαιο των πιθανοτήτων ,για τις τυχαίες μεταβλητές και την από κοινού πιθανότητα, θεωρητικά μπορεί να χρησιμοποιηθεί σε κάθε εφαρμογή των πιθανοτήτων, στην περίπτωση του Μπεϋζιανού συμπερασμού δεν είναι εύκολο κάποιος να κατανοήσει πως θα τα εφαρμόσει. Για αυτόν το λόγο θα χρειαστεί να βρούμε έναν εναλλακτικό ορισμό που θα κάνει πιο εύκολη την εφαρμογή τυχαίων μεταβλητών και από κοινού πιθανοτήτων στον Μπεϋζιανό συμπερασμό.

Στον Μπεϋζιανό συμπερασμό πάντα υπάρχει μια οντότητα που έχει ορισμένα χαρακτηριστικά, την κατάσταση των οποίων θέλουμε να καθορίσουμε που όμως δεν μπορούμε να καθορίσουμε με σιγουριά. Επομένως προσπαθούμε να προσδιορίσουμε πόσο πιθανό είναι ένα συγκεκριμένο χαρακτηριστικό να βρίσκεται σε μια συγκεκριμένη κατάσταση. Μια οντότητα μπορεί να είναι ένα μεμονωμένο σύστημα ή ένα σύνολο συστημάτων. Ένα παράδειγμα μεμονωμένου συστήματος είναι ένας νέος πελάτης μιας τράπεζας, για τον οποίο θα θέλαμε να προσδιορίσουμε με βάση τα οικονομικά στοιχεία του αν ωφελεί την τράπεζα να τον εντάξει στους πελάτες της. Ενώ για τα σύνολα οντοτήτων ένα παράδειγμα θα μπορούσε να είναι τα οικονομικά στοιχεία που έχει η τράπεζα για ένα σύνολο πελατών, και σε αυτή τη περίπτωση το ζητούμενο μπορεί να είναι να βρεθούν πιθανές απάτες.

Στην περίπτωση που μια τυχαία μεταβλητή αναπαριστά ένα χαρακτηριστικό μιας οντότητας που μοντελοποιείται και δεν είμαστε σίγουροι για τις τιμές ή αλλιώς καταστάσεις του χαρακτηριστικού αυτής της συγκεκριμένης οντότητας. Για να ξεπεραστεί αυτό το εμπόδιο αναπτύσσουμε πιθανοτικές σχέσεις ανάμεσα στις μεταβλητές. Όταν μιλάμε για σύνολα οντοτήτων υποθέτουμε πως οι οντότητες του συνόλου έχουν όλες τις ίδιες πιθανοτικές σχέσεις για τις μεταβλητές που χρησιμοποιούνται μέσα στο μοντέλο, χωρίς αυτή τη παραδοχή η ανάλυση σε Μπεϋζιανό επίπεδο δεν είναι εφικτή.

Για να μοντελοποιήσουμε προβλήματα όπως αυτά που προαναφέρθηκαν μπορούμε να ορίσουμε μια τυχαία μεταβλητή X ως το σύμβολο που αντιπροσωπεύει οποιοδήποτε σύνολο τιμών και

ονομάζεται χώρος του X. Για λόγους απλότητας, υποθέτουμε πως ο χώρος του X είναι μετρήσιμος αλλά η συγκεκριμένη θεώρηση επεκτείνεται και στην περίπτωση που ο χώρος δεν θα ήταν μετρήσιμος.

Έστω ένα σύνολο n τυχαίων μεταβλητών $V = \{X_1, X_2, X_3, \ldots, X_n\}$. Η συνάρτηση με την οποία αναθέτουμε έναν πραγματικό αριθμό $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ σε οποιοδήποτε συνδυασμό τιμών του $x_ι$ τέτοιο ώστε η τιμή του x_i να επιλέγεται από το χώρο του X_i, ονομάζεται από κοινού κατανομή πιθανοτήτων μιας τυχαίας μεταβλητής του V αν και μόνο αν ικανοποιεί τις εξής συνθήκες [8]:

1. Για κάθε συνδυασμό τιμών του x_i ισχύει πως : $0 \leq P(X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n) \leq 1$.
2. Έχουμε πως : $\sum_{x_1, x_2, \ldots x_n} P(X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n) = 1$.

6.3 Συνθήκη Μαρκόφ – Markov condition

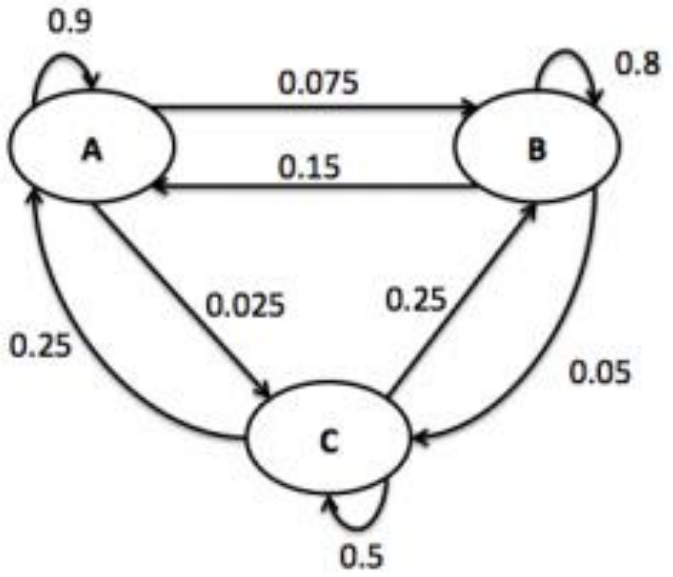
Ο Μπεϋζιανός συμπερασμός είναι αρκετά απλός όταν έχουμε μόνο δύο σχετιζόμενες μεταβλητές, αλλά είναι αρκετά πολύπλοκος όταν προσπαθούμε να εξάγουμε συμπεράσματα έχοντας πολλές σχετιζόμενες μεταβλητές.

Έστω P μια από κοινού πιθανοτική κατανομή P κάποιων τυχαίων μεταβλητών από ένα σύνολο V και έστω ένα άκυκλο γράφημα G = (V,E). Θα λέμε πως το (G,P) ικανοποιεί την συνθήκη Μαρκόφ αν για κάθε μεταβλητή $X \in V$, όπου το {X} είναι υπό όρους ανεξάρτητο με το σύνολο όλων των μη απογόνων δεδομένου του συνόλου όλων των γονιών[8]. Δηλαδή έστω PA_X το σύνολο των γονιών του X και ND_X το σύνολο των μη απογόνων, τότε έχουμε $I_P(\{X\}, ND_X | PA_X)$.

Αν X είναι η ρίζα τότε το σύνολο των γονέων PA_X είναι κενό, επομένως σε αυτή τη περίπτωση η συνθήκη Μαρκόφ σημαίνει πω το {X} είναι ανεξάρτητο του ND_X, δηλαδή $I_P(\{X\}, ND_X)$. Όταν για ένα άκυκλο γράφημα G και μια πιθανοτική κατανομή P το (G,P) ικανοποιεί την συνθήκη Μαρκόφ, τότε το P είναι ίσο με το γινόμενο των υπό όρους πιθανοτήτων κάθε κόμβου που βρίσκονται στο σύνολο των γονέων και ανήκουν στο G.

6.4 Αλυσίδες Μαρκόφ – Markov Chains

Μια αλυσίδα Μαρκόφ μπορεί να οριστεί από ένα σύνολο καταστάσεων Val(X) και ένα μοντέλο που ορίζει για κάθε κατάσταση $x \in Val(X)$ μια κατανομή για την επόμενη κατάσταση πάνω στο Val(X). Πιο συγκεκριμένα, ένα μοντέλο μεταβάσεων «τ» καθορίζει για κάθε ζευγάρι καταστάσεων x,x' η πιθανότητα $\tau(x \rightarrow x')$ να συμβεί μετάβαση από το x στο x'.



Εικόνα 10 Παράδειγμα μιας αλυσίδας Μαρκόφ

Ας πάρουμε ως παράδειγμα την παραπάνω εικόνα που παρουσιάζει μια αλυσίδα Μαρκόφ. Σε μια αλυσίδα Μαρκόφ πρέπει όλες οι μεταβάσεις, που παρουσιάζονται ως ακμές, ενός κόμβου να έχουν άθροισμα ένα (1). Ο πίνακας μεταβάσεων της συγκεκριμένης αλυσίδας Μαρκόφ είναι :

$$P_{transition} = \begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

Στον παραπάνω πίνακα κάθε γραμμή είναι ο κόμβος από τον οποίο γίνεται μετάβαση και κάθε στήλη είναι ο κόμβος στον οποίο γίνεται η μετάβαση. Για παράδειγμα η πρώτη γραμμή είναι όλες οι μεταβάσεις που γίνονται από τον κόμβο A προς τον εαυτό του και τους άλλους δύο κόμβους, παρατηρούμε επίσης πως όντως αν αθροίσουμε τους αριθμούς το άθροισμα είναι ίσο με 1.

Έστω πως στη στιγμή n ξέρουμε πως η κατάσταση του B είναι ίση με ένα, δηλαδή B = 1. Για να υπολογίσουμε τις αλλαγές που γίνονται στο σύστημα για n+1 θα πρέπει να πολλαπλασιάσουμε το πίνακα μεταβάσεων με την τωρινή κατάσταση. Αφού ξέρουμε πως το B είναι η τωρινή κατάσταση έχουμε ένα διάνυσμα [0 1 0].

$$[0\ 1\ 0]\begin{bmatrix} 0.9 & 0.075 & 0.025 \\ 0.15 & 0.8 & 0.05 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} = [0.15\ 0.8\ 0.05]$$

Οι αριθμοί αυτοί μας λένε πως για να μεταβούμε από το B στο A η πιθανότητα είναι 0.15, η πιθανότητα για μετάβαση από το B στο B είναι 0.8 και τέλος η πιθανότητα της μετάβασης από το B στο C είναι 0.05. Αν κάποιος ήθελε να υπολογίσει τη πιθανότητα της ακολουθίας $A \rightarrow B \rightarrow B \rightarrow C$ θα έπρεπε να υπολογίσουμε τη πιθανότητα κάθε μετάβασης και μετά να τις πολλαπλασιάσουμε όλες μαζί. Δηλαδή :

$$P(A \rightarrow B \rightarrow B \rightarrow C) = P(A \rightarrow B)P(B \rightarrow B)P(B \rightarrow C) = 0.075 * 0.8 * 0.05 = 0.003$$

6.5 Μπεϋζιανά Δίκτυα – Bayesian Networks

Τα Μπεϋζιανά δίκτυα διευθετούν τα προβλήματα της αναπαράστασης κατανομών από κοινού πιθανοτήτων για μεγάλο αριθμό τυχαίων μεταβλητών και της χρήσης Μπεϋζιανού συμπερασμού με αυτές τις τυχαίες μεταβλητές.

Ένα Μπεϋζιανό δίκτυο είναι ένα κατευθυνόμενο άκυκλο γράφημα όπου σε κάθε κόμβο παρουσιάζεται μια τυχαία μεταβλητή και κάθε ακμή αντιπροσωπεύει μια άμεση σχέση από τον κόμβο πηγή προς τον κόμβο προορισμό. Κάθε μεταβλητή – κόμβος είναι ανεξάρτητη από το σύνολο των κόμβων που δεν αποτελούν απογόνους της, υπό τη συνθήκη ότι είναι δεδομένο το σύνολο των κόμβων που είναι γονείς της. Τα δίκτυα αυτά είναι γνωστά και ως δίκτυα πεποίθησης (belief networks). Έστω X ένα σύνολο τυχαίων μεταβλητών, η πιθανότητα ενός Μπεϋζιανού δικτύου [9] ορίζεται ως :

$$P_c(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

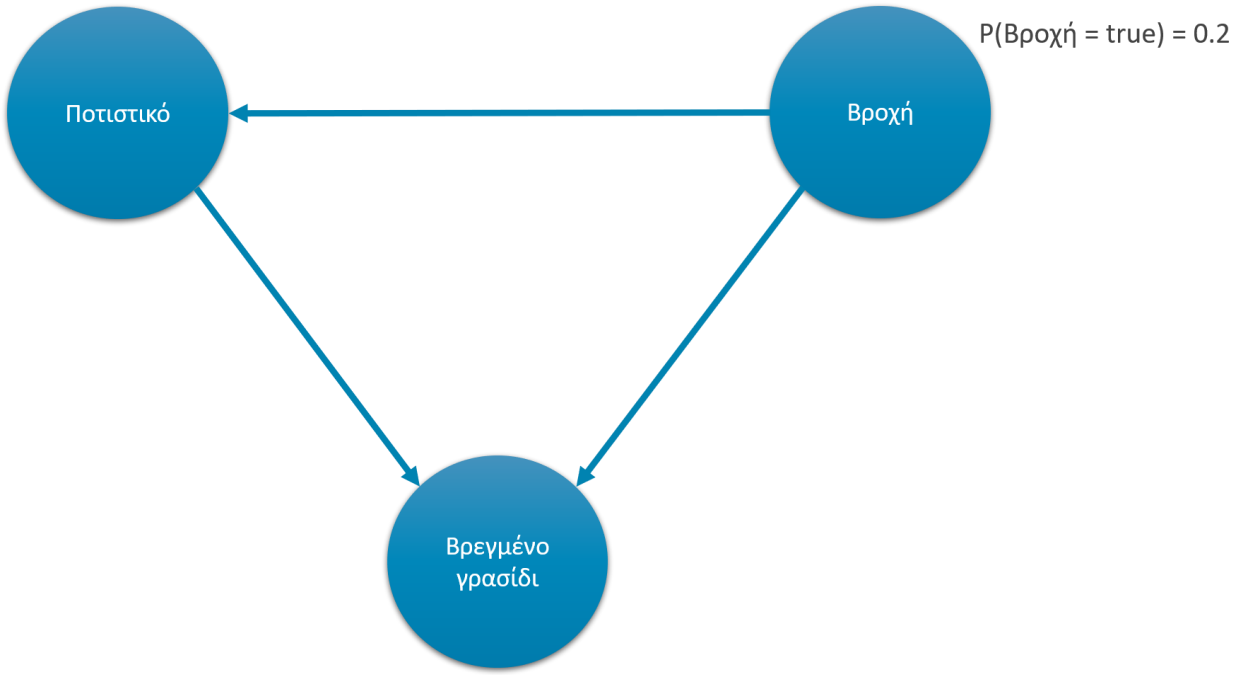
Στην παραπάνω εξίσωση στηρίζεται και η φόρμουλα για τον υπολογισμό συμπερασμάτων μέσω των Μπεϋζιανών δικτύων. Έστω e ένα σύνολο γνωστών μεταβλητών του δικτύου και έστω Y ένα σύνολο μεταβλητών του δικτύου που δεν ξέρουμε τη κατάστασή τους. Για να βρούμε τη τιμή μιας μεταβλητής X λύνουμε την εξίσωση :

$$P(X|e) = a\,P(X,e) = a\left[\sum_{y\in Y} P(X,e,y)\right]$$

$$\text{όπου } a = \frac{1}{\sum_{x\in X} P(X=x,e)}$$

Το α είναι ένας συντελεστής κανονικοποίησης για το P(X|e).

Ας δούμε όμως ένα κλασικό παράδειγμα εφαρμογής ενός Μπεϋζιανού δικτύου. Έστω ένα Μπεϋζιανό δίκτυο όπως αυτό παρουσιάζεται στην εικόνα και με τους παρακάτω πίνακες πιθανοτήτων.



Βροχή	P(Ποτιστικό = true)
True	0.01
False	0.4

Ποτιστικό	Βροχή	P(Βρεγμένο γρασίδι = true)
True	True	0.99
True	False	0.90
False	True	0.80
False	False	0.00

$P(B = true\,|\Gamma = true) = a\,P(B = true)\ \times \sum_{(\pi)\epsilon\Pi} P(\Pi = \pi\,|B = true)\ \times P(\Gamma = true|\Pi = \pi, B = true)$

$=>\ (B = true\,|\Gamma = true) = a\ 0.2\ \times$
$[P(\Pi = true|B = true)P(\Gamma = true|\Pi = true, B = true) +$
$P(\Pi = false|B = true)P(\Gamma = true|\Pi = false, B = true)]$

$=> P(B = true|\Gamma = true) = a\ 0.2\ \times [0.01 \times 0.99 + 0.99\ \times 0.8] = a\ 0.1604$

Τέλος κανονικοποιούμε το αποτέλεσμα μας με τον συντελεστή α, ο οποίος υπολογίζεται από το εξής κλάσμα :

$$a = \frac{1}{P(B = true|\Gamma = true) + P(B = false|\Gamma = true)}$$

Για να λύσουμε την συγκεκριμένη εξίσωση θα πρέπει να βρούμε την πιθανότητα το γρασίδι να είναι βρεγμένο ενώ δεν έχει βρέξει.

$P(B = false|\Gamma = true) = a\,P(B = false) \times \sum_{(\pi)\epsilon\Pi} P(\Pi = \pi\,|B = false)\ \times P(\Gamma = true|\Pi = \pi, B = false)$

$=>\ (B = false\,|\Gamma = true)$
$= a\ 0.8\ \times [P(\Pi = true|B = false)P(\Gamma = true|\Pi = true, B = false)$
$+ P(\Pi = false|B = false)P(\Gamma = true|\Pi = false, B = false)]$

$=> P(B = false|\Gamma = true) = a\ 0.8\ \times [0.4 \times 0.9 + 0.6 \times 0] = a\ 0.288$

Άρα το α είναι :

$$\alpha = \frac{1}{0.1604 + 0.288} = \frac{1}{0.4484}$$

Έχοντας τον συντελεστή α οι τελικές πιθανότητες υπολογίζονται ως εξής :

$$P(B = true\,|\Gamma = true) = a\,0.1604 = 0.3577$$

και

$$P(B = false\,|\Gamma = true) = a\,0.288 = 0.6423$$

6.6 Μάθηση δομής – Structure learning

Συχνά το πιο δύσκολα διαχειρίσιμο μέρος των Μπεϋζιανών δικτύων είναι το τι κάνουμε όταν δεν έχουμε ένα γραφικό μοντέλο που να περιγράφει τη δομή των δεδομένων μας, δηλαδή η μάθηση της δομής του συγκεκριμένου δικτύου. Η μάθηση δομής είναι ένα σύνολο μεθόδων μέσω των οποίων βρίσκουμε ένα κατευθυνόμενο άκυκλο γράφημα το οποίο αναπαριστά την δομή των σχέσεων ενός συνόλου δεδομένων με n μεταβλητές. Μπορούμε να χωρίσουμε τους διάφορους αλγορίθμους μάθησης δομής σε τρεις κατηγορίες [11]:

1. τους constraint-based αλγορίθμους, που χρησιμοποιώντας στατιστικές δοκιμές βρίσκουμε περιορισμούς όσον αναφορά τις υπό όρους εξαρτήσεις των κόμβων μας και στη συνέχεια συνδέουμε κόμβους που δεν είναι ανεξάρτητοι.
2. τους score-based αλγορίθμους, όπου σε κάθε υποψήφιο κατευθυνόμενο άκυκλο γράφημα αναθέτεται ένα σκορ με βάση κάποια μέθοδο αξιολόγησης (scoring method).
3. Τους υβριδικούς αλγορίθμους οι οποίοι συνδυάζουν τις δυο παραπάνω κατηγορίες καθώς πρώτα χρησιμοποιούν κάποια constraint-based προσέγγιση για να μειώσουν τον χώρο των πιθανών κατευθυνόμενων άκυκλων γραφημάτων και στη συνέχεια χρησιμοποιούν μια στρατηγική score-based για να καταλήξουν στο βέλτιστο γράφημα.

7 Προτεινόμενη προσέγγιση

Το πρακτικό μέρος της παρούσας εργασίας χωρίζεται σε δύο διαφορετικά μέρη. Στο πρώτο μέρος θα μελετηθούν κλασικοί αλγόριθμοι εξόρυξης διεργασιών (Alpha miner, Heuristic miner και Inductive miner) με τους οποίους θα εξάγουμε μοντέλα διαδικασιών από ένα αρχείο καταγραφής γεγονότων. Τα αποτελέσματα των αλγορίθμων θα καταγραφούν και στη συνέχεια θα αξιολογηθούν μέσω διαφόρων τεχνικών (Fitness,Precision,Generalization και Simplicity). Στο δεύτερο μέρος ξεκινώντας με το ίδιο αρχείο καταγραφής γεγονότων προσπαθούμε να δημιουργήσουμε ένα Μπεϋζιανό δίκτυο από το οποίο θα βγάλουμε κάποια συμπεράσματα. Επειδή τα Μπεϋζιανά δίκτυα είναι ένα σχετικά νέο εργαλείο στην φαρέτρα της εξόρυξης διεργασιών δεν υπάρχει κάποιος συγκεκριμένος αλγόριθμος για να γίνει εξαγωγή ενός τέτοιου δικτύου από ένα αρχείο καταγραφής γεγονότων. Για τον λόγο αυτό στην εργασία γίνεται μια προσπάθεια να χρησιμοποιηθούν Μπεϋζιανά δίκτυα στην εξαγωγή διεργασιών και να αξιολογηθεί κατά πόσο αυτό είναι εφικτό, με ποιον τρόπο και ποια είναι τα αποτελέσματα.

7.1 Μέρος 1ο – Κλασικοί Αλγόριθμοι

Οι αλγόριθμοι εξόρυξης διεργασιών έχουν ως στόχο την εύρεση ενός μοντέλου διεργασιών που περιγράφει τις σχέσεις των δραστηριοτήτων (activities) που εκτελούνται κατά την διάρκεια εκτέλεσης της διεργασίας. Οι αλγόριθμοι που θα μελετηθούν είναι ο alpha miner, ο heuristic miner και τέλος ο inductive miner. Αφού γίνει η εξαγωγή των μοντέλων θα αξιολογήσουμε τον βαθμό στον οποίο το μοντέλο περιγράφει σωστά το αρχείο καταγραφής γεγονότων. Η αξιολόγηση αυτή θα γίνει σε τέσσερις τομείς replay fitness, precision, generalization, simplicity. Η υλοποίηση των παραπάνω γίνεται με την χρήση της python και της βιβλιοθήκης pm4py.

Στόχος του replay fitness είναι να υπολογίσει αν η συμπεριφορά του αρχείου καταγραφής γεγονότων καθρεπτίζεται στο μοντέλο διεργασιών. Υπάρχουν διάφορες μέθοδοι για τον υπολογισμό του replay fitness, δυο από αυτές είναι η token-based και η alignments. Για το token-based replay υπολογίζεται το ποσοστό των traces που περιγράφουν επακριβώς τα δεδομένα του αρχείου καταγραφής γεγονότων, καθώς και μια τιμή που δηλώνει το fitness του μοντέλου. Η διαφοροποίηση του alignment είναι στην τιμή που επιστρέψει καθώς υπολογίζει για κάθε trace μια τιμή fitness και τελικά επιστρέφει τον μέσο όρο αυτών των τιμών ως συνολικό fitness του μοντέλου.

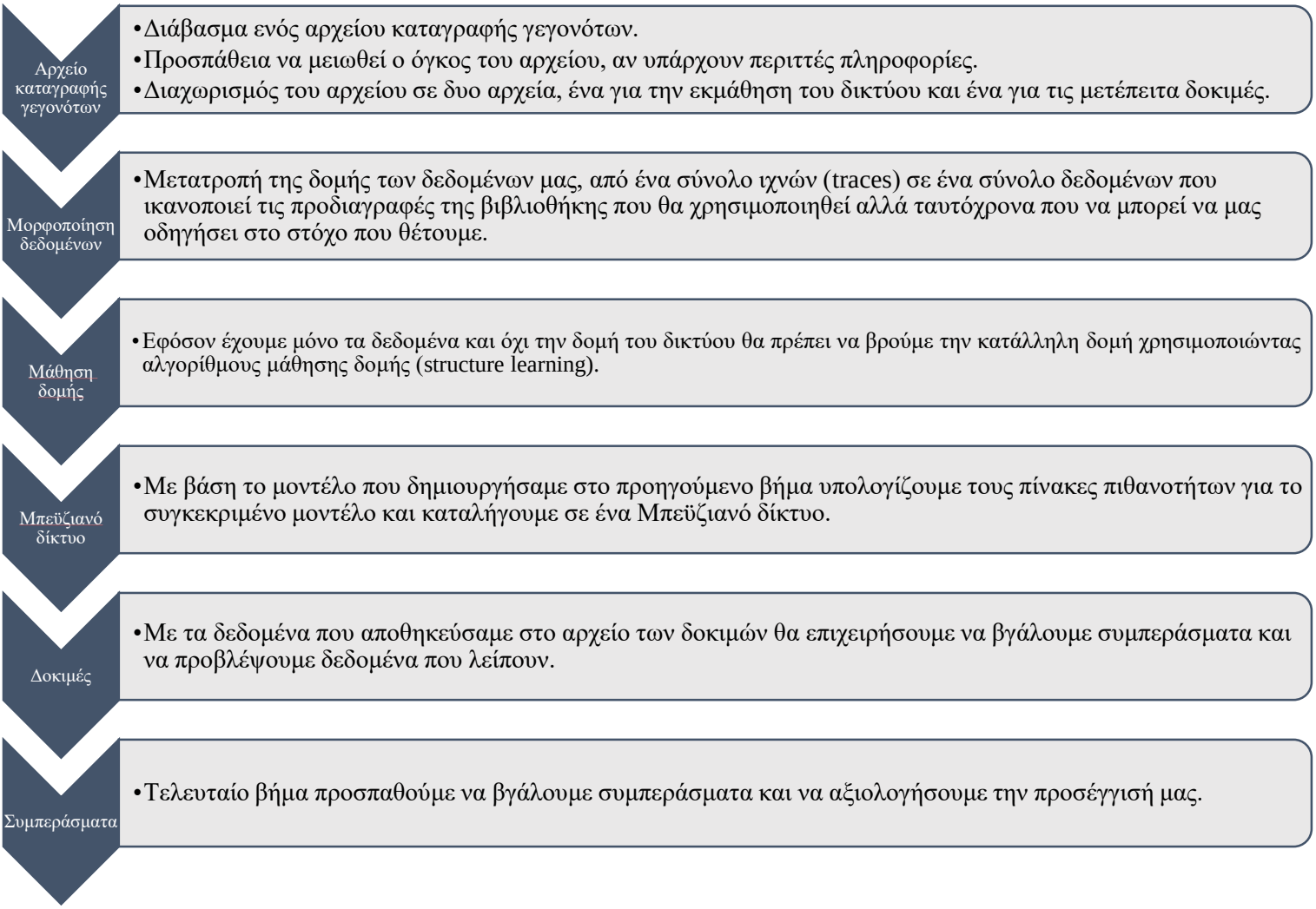
Το μοντέλο που δημιουργήθηκε δεν θα πρέπει να απεικονίζει κάποια συμπεριφορά η οποία να μην υπάρχει στα δεδομένα, για αυτό το λόγο υπάρχει και ο έλεγχος του Precision. Στόχος του ελέγχου του precision είναι να ποσοτικοποιήσει το ποσοστό των «συμπεριφορών» που επιτρέπει το μοντέλο διεργασιών να εκτελεστούν οι οποίες δεν ανήκουν στο αρχείο καταγραφής γεγονότων.

Επιπλέον το μοντέλο πρέπει να γενικεύει την συμπεριφορά που υπάρχει στο αρχείο καταγραφής, έτσι ώστε να μην υπάρχει πολύπλοκη δομή και να μην στοχεύει μόνο στα συγκεκριμένα δεδομένα που του δόθηκαν αλλά να μπορεί να αναπαραστήσει γενικά την διαδικασία.

Τέλος ο έλεγχος για το simplicity του μοντέλου ελέγχει πόσο απλό είναι το μοντέλο. Στόχος είναι το παραχθέν μοντέλο να είναι όσο πιο απλό γίνεται.

7.2 Μέρος 2ο – Εξόρυξη διεργασιών και Bayesian Networks

Οι αλγόριθμοι του πρώτου μέρους χρησιμοποιούνται για την δημιουργία ενός μοντέλου που θα περιγράφει όσο το δυνατόν καλύτερα ένα σύνολο δεδομένων μιας διαδικασίας. Όμως όπως αναλύθηκε στο θεωρητικό μέρος της εργασίας η εξόρυξη διεργασιών έχει και άλλες λειτουργίες εκτός της ανακάλυψης νέων μοντέλων. Πολύ συχνά στις επιχειρήσεις πρέπει να γίνει μια ανάλυση των διαδικασιών για τις οποίες όμως υπάρχει μια αβεβαιότητα. Με τον όρο αβεβαιότητα εννοούμε την προσπάθεια να εκτιμήσουμε την πιθανότητα μιας ακολουθίας εργασιών (tasks) να συμβούν με δεδομένο πως μπορούμε να γνωρίζουμε/παρατηρούμε μόνο ένα μέρος των συνολικών εργασιών που συντελούν την διεργασία. Για αυτόν τον λόγο προτείνουμε την χρήση Μπεϋζιανών δικτύων για να διαχειριστούμε την αβεβαιότητα στην εξόρυξη διεργασιών. Στόχος της συγκεκριμένης εργασίας είναι παρουσιάσει σύστημα που αρχίζοντας από ένα αρχείο καταγραφής γεγονότων να μπορέσει να δημιουργήσει ένα Μπεϋζιανό δίκτυο στο οποίο τελικά θα δοκιμάσουμε να εκτιμήσουμε αν μια δραστηριότητα θα συμβεί.



7.2.1 Βήμα 1ο – Αρχείο Καταγραφής Γεγονότων

Αρχικά έχουμε ένα αρχείο καταγραφής γεγονότων όπου κρατούνται όλα τα στοιχεία κάποιας διεργασίας. Όπως ορίσαμε σε προηγούμενο κεφάλαιο ένα αρχείο καταγραφής γεγονότων (event log) είναι ένα αρχείο στο οποίο αποθηκεύονται με συγκεκριμένο τρόπο όλα τα δεδομένα που αφορούν τις διαδικασίες μια επιχείρησης. Επειδή όμως δεν ξέρουμε αν χρειαζόμαστε όλα αυτά τα στοιχεία πρέπει να μελετήσουμε τι στοιχεία έχουν κρατηθεί στο συγκεκριμένο αρχείο. Ανάλογα με το στόχο μας ή το τι θέλουμε να αναλύσουμε ίσως να μην χρειαστούν όλα αυτά τα δεδομένα. Περιττά στοιχεία πρέπει να διαγραφούν, έτσι ώστε να είναι πιο εύκολα διαχειρίσιμα και η όλη διαδικασία να είναι πιο γρήγορη. Επιπλέον στο βήμα αυτό χωρίζεται το αρχείο καταγραφής γεγονότων σε δυο αρχεία, αφού θα χρειαστούμε ένα μέρος τον δεδομένων για να δημιουργήσουμε το Μπεϋζιανό δίκτυο και ένα μέρος δεδομένων που θα χρησιμοποιήσουμε για

δοκιμές πάνω στο δίκτυο που δημιουργήσαμε. Επομένως από το αρχικό αρχείο δημιουργούμε ένα αρχείο για την μάθηση του δικτύου και ένα για τις δοκιμές, το πρώτο θα περιέχει το 80% των δεδομένων ενώ το αρχείο δοκιμών θα περιέχει το υπόλοιπο 20% του συνόλου των αρχικών δεδομένων.

7.2.2 Βήμα 2ο – Μορφοποίηση Δεδομένων

Τα αρχικά δεδομένα που βρίσκονται στο αρχείο καταγραφής γεγονότων αποτελούνται από ένα σύνολο ιχνών (traces), όπου κάθε ίχνος είναι μια σειρά από δραστηριότητες (activities). Τα δεδομένα αυτά όμως δεν είναι κατάλληλα για να μπορέσουμε να δημιουργήσουμε ένα Μπεϋζιανό δίκτυο. Στο σημείο αυτό πρέπει να αποφασιστεί ποιος είναι ο στόχος που θέλουμε να επιτύχουμε, για παράδειγμα στόχος του δικτύου που θα δημιουργήσουμε σε αυτή την εργασία είναι να μπορούμε να εκτιμήσουμε αν μια δραστηριότητα (activity) θα συμβεί ή όχι, γνωρίζοντας την κατάσταση των υπολοίπων δραστηριοτήτων.

Έτσι λοιπόν αφού αποφασιστεί ο στόχος πρέπει στη συνέχεια να μετατραπούν τα δεδομένα στην κατάλληλη μορφή. Από το αρχείο μάθησης που δημιουργήσαμε στο προηγούμενο βήμα θα πρέπει να εξάγουμε όλες τις μοναδικές δραστηριότητες. Στη συνέχεια για κάθε ίχνος του αρχείου θα πρέπει να ελέγξουμε ποιες δραστηριότητες έχουν συμβεί και ποιες όχι, έτσι δημιουργείται ένας πίνακας με γραμμές ίσες με τον αριθμών των ιχνών και στήλες ίσες με τον αριθμό των δραστηριοτήτων. Ο πίνακας αυτός περιέχει τα δεδομένα πάνω στα οποία θα στηριχτούμε για να δημιουργήσουμε το Μπεϋζιανό δίκτυο. Παρακάτω παραθέτουμε ένα παράδειγμα του τελικού πίνακα των δεδομένων μας.

	A_Accepted	**A_Cancelled**	**A_Complete**	**A_Concept**	**A_Create Application**	**A_Denied**	**A_Incomplete**
0	present	absent	present	present	present	absent	present
1	present	absent	present	present	present	present	absent
2	present	absent	present	present	present	absent	present
3	present	absent	present	present	present	absent	present
4	present	present	present	present	present	absent	absent

7.2.3 Βήμα 3ο – Μάθηση Δομής

Θέλουμε να φτάσουμε σε ένα Μπεϋζιανό δίκτυο, χωρίς να έχουμε όμως το μοντέλο που περιγράφει αυτό το δίκτυο παρά μόνο δεδομένα μέσω των οποίων θα πρέπει να φτάσουμε στο δίκτυο. Για τον λόγο αυτό χρησιμοποιώντας έναν αλγόριθμο μάθησης δομής (structure learning) με την βοήθεια της βιβλιοθήκης pgmpy θα προσπαθήσουμε να βρούμε την δομή που περιγράφει σωστά τα δεδομένα του προηγούμενου βήματος.

Πιο συγκεκριμένα παίρνοντας τα δεδομένα θα προσπαθήσουμε να εξάγουμε το μοντέλο που περιγράφει τα δεδομένα με έναν constrain based αλγόριθμο, τον PC. Ο αλγόριθμος PC χρησιμοποιείται για τον υπολογισμό ενός σκελετού για ένα άκυκλο κατευθυνόμενο γράφημα πολλών μεταβλητών. Πλεονέκτημά του είναι η ταχύτητα που μας παρέχει για προβλήματα που έχουν πολλούς κόμβους, όπως στην περίπτωσή μας όπου κάθε κόμβος είναι μια δραστηριότητα. Στην εικόνα 11 παρουσιάζουμε έναν ψευδοκώδικα του αλγορίθμου PC.

Algorithm 1 The PC$_{pop}$-algorithm

INPUT: Vertex Set V, Conditional Independence Information
OUTPUT: Estimated skeleton C, separation sets S (only needed when directing the skeleton afterwards)
Form the complete undirected graph $\tilde{C}$ on the vertex set V.
$\ell = -1; \quad C = \tilde{C}$
repeat
  $\ell = \ell + 1$
  repeat
    Select a (new) ordered pair of nodes i,j that are adjacent in C such that $|adj(C,i) \setminus \{j\}| \geq \ell$
    repeat
      Choose (new) $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$.
      if i and j are conditionally independent given $\mathbf{k}$ **then**
        Delete edge i,j
        Denote this new graph by C
        Save $\mathbf{k}$ in $S(i,j)$ and $S(j,i)$
      end if
    until edge i,j is deleted or all $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been chosen
  until all ordered pairs of adjacent variables i and j such that $|adj(C,i) \setminus \{j\}| \geq \ell$ and $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been tested for conditional independence
until for each ordered pair of adjacent nodes i,j: $|adj(C,i) \setminus \{j\}| < \ell$.

Εικόνα 11: Ψευδοκώδικας του PC αλγορίθμου [10]

Στόχος της εργασίας δεν είναι η σύγκριση και η αξιολόγηση των διαφόρων αλγορίθμων εκμάθησης δομής και επομένως δεν θα αναλύσουμε λεπτομερώς το συγκεκριμένο κομμάτι. Ο μόνος στόχος μας είναι η δημιουργία του Μπευζιανού δικτύου.

7.2.4 Βήμα 4ο – Μπεϋζιανό Δίκτυο

Αποτέλεσμα του προηγούμενου βήματος είναι η εξαγωγή ενός μοντέλου που αντιπροσωπεύει τα δεδομένα του αρχείου καταγραφής γεγονότων. Το μοντέλο είναι ένα άκυκλο κατευθυνόμενο γράφημα και επομένως το μόνο στοιχείο που λείπει για την δημιουργία του Μπεϋζιανού δικτύου είναι η προσθήκη των πινάκων πιθανοτήτων για κάθε κόμβο του μοντέλου. Στην εργασία αυτή οι κόμβοι του κατευθυνόμενου άκυκλου γραφήματος αναπαριστούν δραστηριότητες της διαδικασίας και οι πίνακες πιθανοτήτων υπολογίζουν την πιθανότητα μια από αυτές τις δραστηριότητες να συμβεί ή όχι. Ο υπολογισμός των πινάκων πιθανοτήτων για κάθε κόμβο του μοντέλου γίνεται με την χρήση της βιβλιοθήκης pgmpy.

7.2.5 Βήμα 5ο – Δοκιμές

Στο πέμπτο βήμα προσπαθούμε να δοκιμάσουμε αν τελικά μπορούμε να χρησιμοποιήσουμε τα Μπεϋζιανά δίκτυα στην εξόρυξη διεργασιών. Στην συγκεκριμένη εργασία αυτό που προσπαθούμε να πετύχουμε είναι να αφαιρέσουμε την αβεβαιότητα από μια διεργασία. Θα πρέπει λοιπόν αν ελέγξουμε αν το Μπεϋζιανό δίκτυο που δημιουργήσαμε είναι ικανό να προβλέψει δεδομένα της διεργασίας γνωρίζοντας μόνο ένα μέρος της κατάστασης μιας εκτελούμενης διεργασίας.

Έχοντας λοιπόν έτοιμο το Μπεϋζιανό δίκτυό μας πλέον πρέπει να δούμε αν μπορεί να προβλέψει δεδομένα που λείπουν. Παίρνουμε δηλαδή το αρχείο δοκιμών και αφαιρούμε ένα μέρος των δεδομένων, έτσι ώστε να δώσουμε το αρχείο στο δίκτυο και να δούμε αν μπορεί να προβλέψει σωστά τα δεδομένα που αφαιρέσαμε.

Αφού τελειώσει η διαδικασία βλέπουμε το ποσοστό των δεδομένων που το δίκτυο προέβλεψε σωστά καταγράφοντας τα false positives και false negatives, αν αυτά υπάρχουν.

7.2.6 Συμπεράσματα

Τελευταίο βήμα είναι να αξιολογήσουμε αν μπορεί η χρήση των Μπεϋζιανών δικτύων να αντιμετωπίσει την αβεβαιότητα που υπάρχει στον τομέα της εξόρυξης διεργασιών. Καθώς και να βγάλουμε συμπεράσματα από την προσέγγιση που χρησιμοποιήσαμε.

8 Πρακτικό Μέρος

Για να δοκιμάσουμε την προσέγγιση που αναλύσαμε στο προηγούμενο κεφάλαιο, σε αυτό το κεφάλαιο θα πάρουμε ένα αρχείο καταγραφής γεγονότων και θα προσπαθήσουμε να ακολουθήσουμε τα βήματα της προτεινόμενης προσέγγισης. Το κεφάλαιο χωρίζεται σε δύο μέρη όπως και η προσέγγιση, αρχικά θα δοκιμάσουμε παραδοσιακούς αλγορίθμους εξόρυξης διεργασιών στο αρχείο καταγραφής γεγονότων μας, ενώ στη συνέχεια θα προσπαθήσουμε να υλοποιήσουμε ένα Μπεϋζιανό δίκτυο από το ίδιο αρχείο.

Το αρχείο καταγραφής γεγονότων που επιλέχθηκε περιέχει τραπεζικά δεδομένα από διαδικασίες δανείου. Οι δραστηριότητες (activities) που αποτελούν το συγκεκριμένο αρχείο είναι οι εξής :

Δραστηριότητες
A_Accepted
A_Cancelled
A_Complete
A_Concept
A_Create Application
A_Denied
A_Incomplete
A_Pending
A_Submitted
A_Validating
O_Accepted
O_Cancelled
O_Create Offer
O_Created
O_Refused
O_Returned
O_Sent (mail and online)
O_Sent (online only)

W_Assess potential fraud
W_Call after offers
W_Call incomplete files
W_Complete application
W_Handle leads
W_Personal Loan collection
W_Shortened completion
W_Validate application

Η υλοποίηση γίνεται με τη χρήση της Python. Επίσης για την διαχείριση και ανάλυση των δεδομένων χρησιμοποιούμε την βιβλιοθήκη pandas, μια βιβλιοθήκη λογισμικού για την Python για χειρισμό και ανάλυση δεδομένων που προσφέρει δομές δεδομένων όπως τα dataframes. Για το πρώτο μέρος γίνεται επίσης η χρήση της βιβλιοθήκης PM4Py που δημιουργήθηκε για να διευκολύνει στην εξόρυξη διεργασιών που μέχρι τώρα δεν υπήρχε κάποιο συγκεκριμένο λογισμικό ή βιβλιοθήκη για τον κλάδο αυτό, αφού η βιβλιοθήκη προσφέρει επεκτασιμότητα και ευχρηστία, ενώ μπορεί να ενσωματώσει και τη βιβλιοθήκη pandas που αναφέραμε πιο πριν. Η βιβλιοθήκη διαθέτει μεθόδους για την εξόρυξη διεργασιών από ένα σύνολο δεδομένων ή από ένα αρχείο καταγραφής γεγονότων με αλγορίθμους όπως ο alpha miner, ο inductive miner και ο heuristic miner. Για το δεύτερο μέρος για την υλοποίηση του Μπεϋζιανού δικτύου χρησιμοποιείται η βιβλιοθήκη pgmpy, μια βιβλιοθήκη της python για υλοποίηση πιθανοτικών γραφικών μοντέλων. Με τη βοήθεια της συγκεκριμένης βιβλιοθήκης μπορούμε από τα δεδομένα του αρχείου καταγραφής γεγονότων να μάθουμε τη δομή, να υπολογίσουμε τους πίνακες πιθανοτήτων των κόμβων και στη συνέχεια να δημιουργήσουμε ένα Μπεϋζιανό δίκτυο. Για να δοκιμάσουμε τις δυνατότητες του δικτύου θα γίνει πάλι χρήση της ίδια βιβλιοθήκης καθώς θα προσπαθήσουμε να προβλέψουμε την κατάσταση κόμβων για τους οποίους δεν ξέρουμε τη κατάστασή τους.

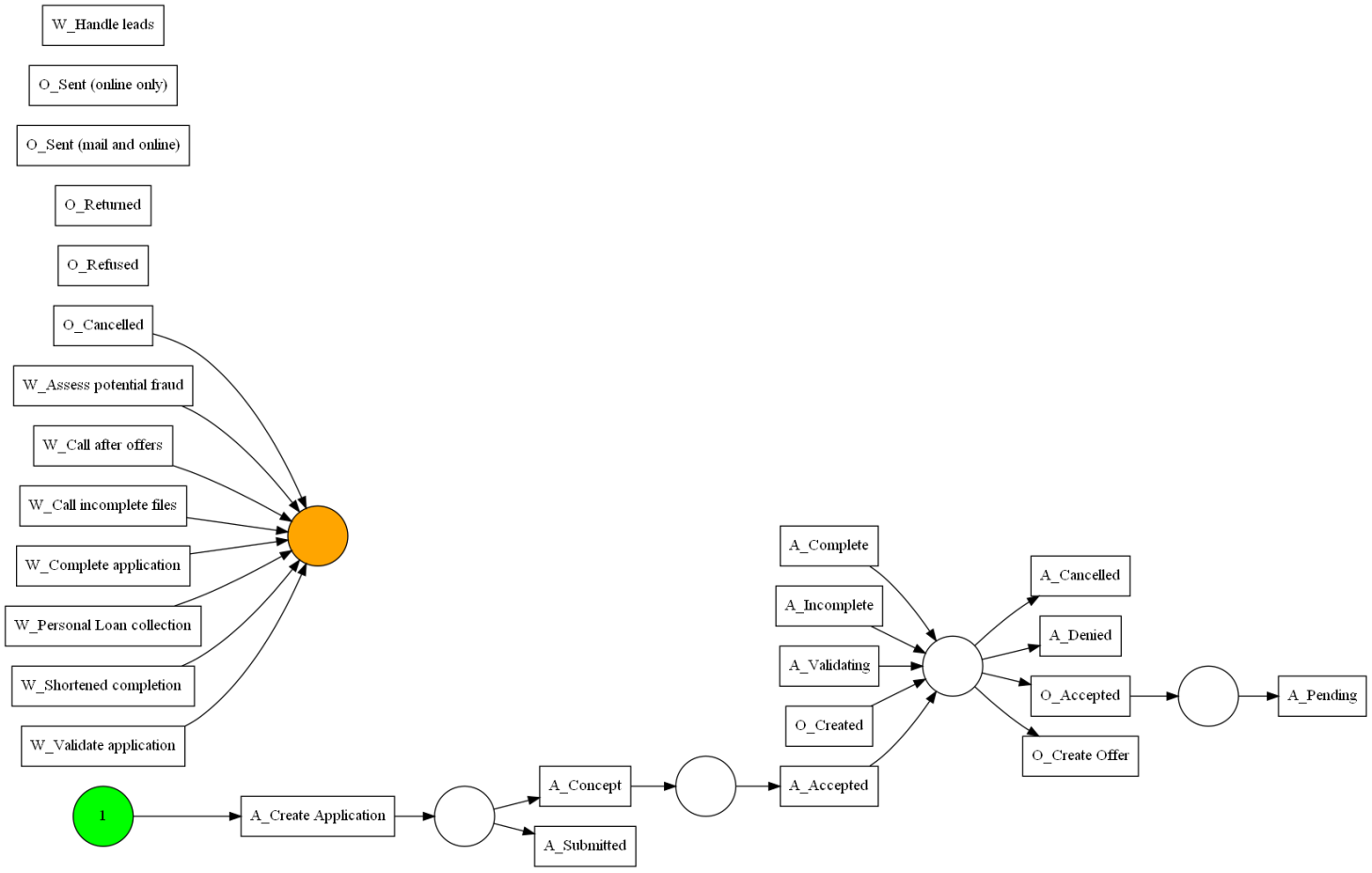
8.1 Πρώτο μέρος – Κλασικοί αλγόριθμοι

8.1.1 Alpha Miner

Όπως είπαμε παραπάνω στο μέρος αυτό θα προσπαθήσουμε να εξάγουμε γραφικά μοντέλα χρησιμοποιώντας αλγορίθμους εξόρυξης διεργασιών. Ο πρώτος αλγόριθμος που θα

χρησιμοποιήσουμε και ένας από τους πιο γνωστούς αλγορίθμους ανακάλυψης διεργασιών είναι ο alpha miner, ο συγκεκριμένος αλγόριθμος μπορεί να βρει :

1. Ένα μοντέλο Petri net όπου όλες οι μεταβάσεις είναι εμφανείς και μοναδικές και αντιστοιχούν στις δραστηριότητες του αρχείου.
2. Ένα αρχικό σύνολο που περιγράφει τη κατάσταση του μοντέλου Petri net όταν ξεκινά η εκτέλεση.
3. Ένα τελικό σύνολο που περιγράφει τη κατάσταση του μοντέλου Petri net όταν τελειώνει η εκτέλεση.



Εικόνα 12: Αποτέλεσμα Alpha miner

Ένα από τα χαρακτηριστικά προβλήματα του Alpha miner είναι πως δεν μπορεί να χειριστεί κύκλους μεγέθους ενός ή δυο κόμβων, επιπλέον το τελικό μοντέλο μπορεί να μην είναι υγιές καθώς ο αλγόριθμος είναι αδύναμος στο θόρυβο. Στην εικόνα 12 βλέπουμε ένα αποτέλεσμα που δεν μας ικανοποιεί ιδιαίτερα καθώς έχουμε ένα ασύνδετο μοντέλο.

Αποτελέσματα Fitness score του αλγορίθμου Alpha Miner		
perc_fit_traces	**average_trace_fitness**	**log_fitness**
0.0	0.5842	0.5638

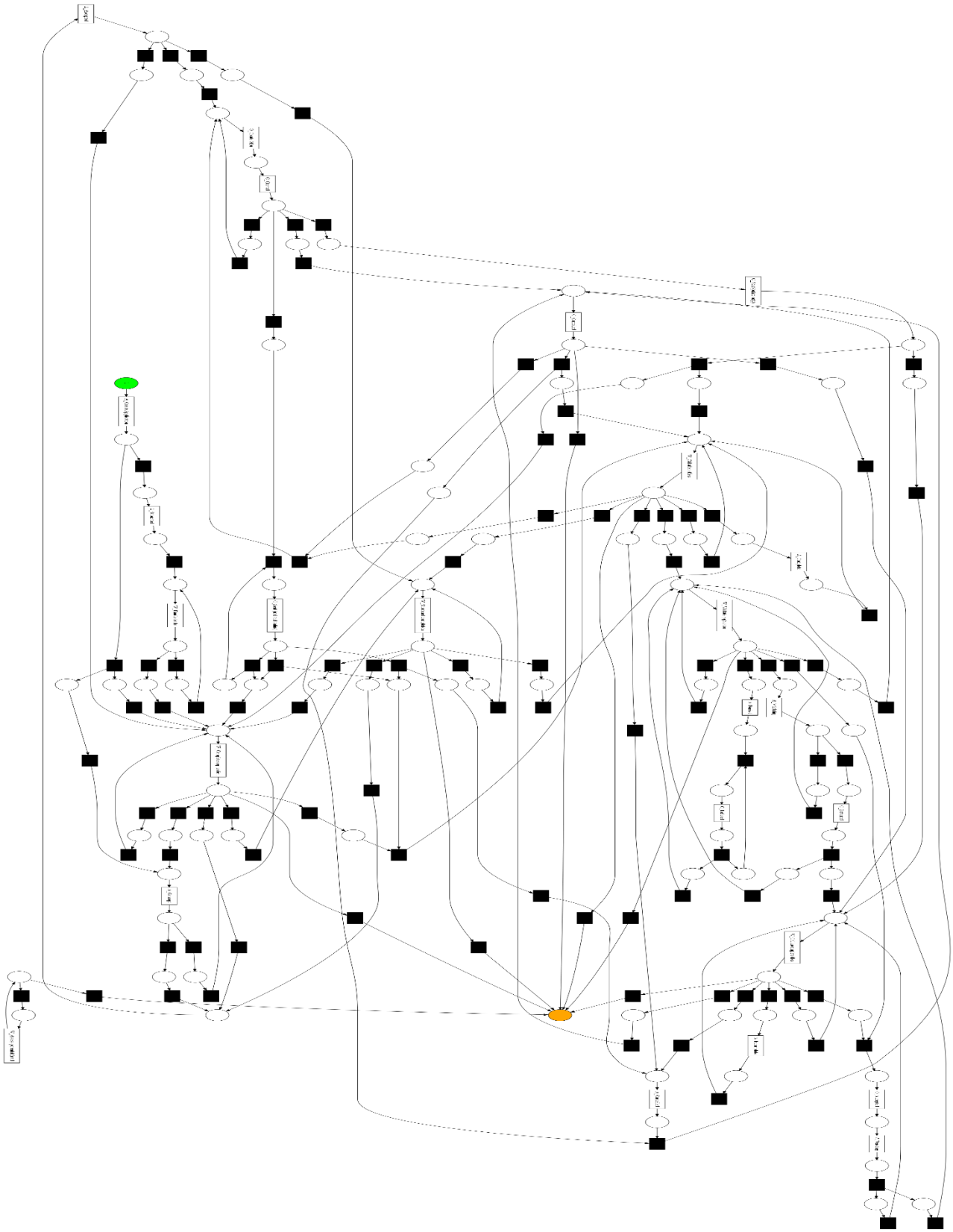
Κοιτώντας τα αποτελέσματα του fitness score βλέπουμε πως το μοντέλο που δημιούργησε ο αλγόριθμος δεν μπορεί να περιγράψει ικανοποιητικά την συμπεριφορά των δεδομένων που υπάρχουν στο αρχείο καταγραφής γεγονότων.

Αποτελέσματα Precision, Generalization και Simplicity για τον Alpha Miner		
Precision	**Generalization**	**Simplicity**
0.0957	0.9815	1.0

Από τα αποτελέσματα του παραπάνω πίνακα θα σταθούμε κυρίως στο σκορ του precision η τιμή του οποίου είναι λίγο μικρότερη από 0.01 και επομένως επαληθεύουμε πως το μοντέλο που παρήγαγε ο Alpha Miner δεν αντιπροσωπεύει τα δεδομένα του αρχείου καταγραφής γεγονότων.

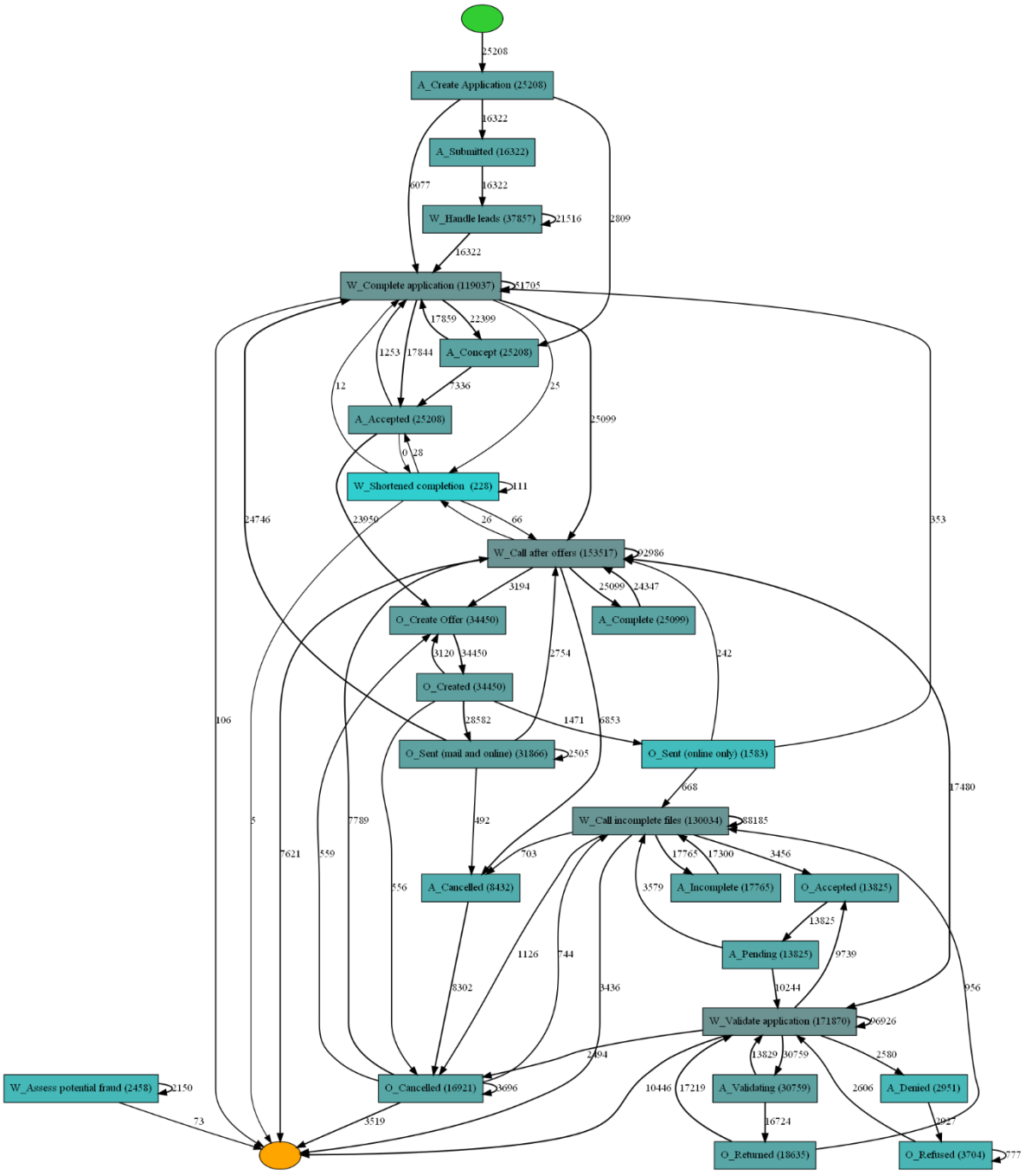
8.1.2 Heuristic Miner

Ο επόμενος αλγόριθμος με τον οποίο θα προσπαθήσουμε να εξάγουμε ένα γραφικό μοντέλο που θα αντιπροσωπεύει τα δεδομένα μας είναι ο Heuristic miner. Ο συγκεκριμένος αλγόριθμος λαμβάνει υπόψη τη συχνότητα κατά τη δημιουργία του μοντέλου. Η ιδέα πίσω από αυτό είναι πως γεγονότα που δεν έχουν μεγάλη συχνότητα εμφάνισης δεν λαμβάνονται υπόψιν κατά την κατασκευή του μοντέλου.



Εικόνα 13 : Petri Net από τον Heuristic αλγόριθμο

Στην εικόνα 13 βλέπουμε το Petri net που δημιουργήθηκε, το οποίο όπως βλέπουμε είναι πολύ πολύπλοκο. Για αυτόν το λόγο εκτός του Petri net χρησιμοποιήσαμε τον heuristic αλγόριθμο για να κατασκευάσουμε και ένα process tree (Εικόνα 14) από τα ίδια δεδομένα με σκοπό να έχουμε ένα πιο ευανάγνωστο αποτέλεσμα.



Εικόνα 14: Process Tree του Heuristic αλγορίθμου

Αποτελέσματα Fitness score του αλγορίθμου Heuristic Miner		
perc_fit_traces	**average_trace_fitness**	**log_fitness**
0.0	0.9642	0.9691

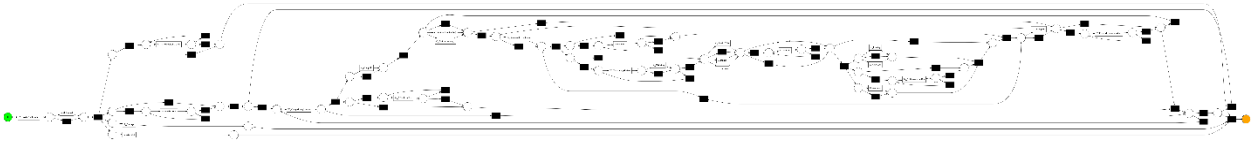
Από τον παραπάνω πίνακα μπορούμε να παρατηρήσουμε ένα εμφανώς καλύτερο αποτέλεσμα. Η τιμή του fitness score είναι κοντά στο 0.96 και επομένως μπορούμε να θεωρήσουμε πως το συγκεκριμένο γραφικό μοντέλο περιγράφει ικανοποιητικά το αρχείο καταγραφής γεγονότων που του εισάγαμε.

Αποτελέσματα Precision, Generalization και Simplicity για τον Heuristic Miner		
Precision	**Generalization**	**Simplicity**
0.6246	0.9605	0.7048

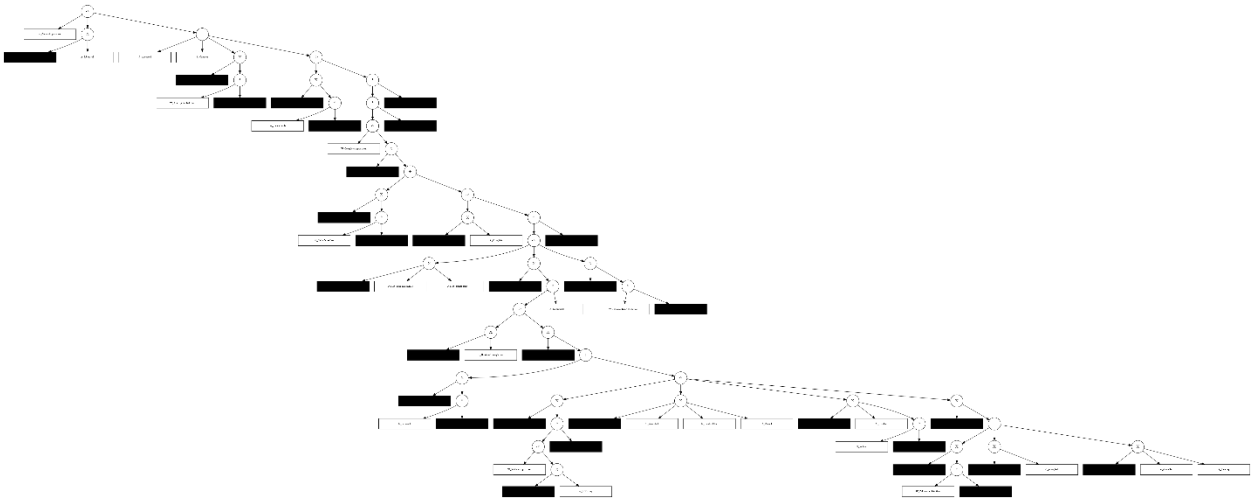
Στα αποτελέσματα των Precision, Generalization και Simplicity scores αξίζει να σημειώσουμε πως ο heuristic miner έχει precision κοντά στο 0.62 μια τεράστια αύξηση σε σχέση με το αποτέλεσμα του alpha miner όπου το score αυτό ήταν κοντά στο 0.01. Επιπλέον όμως παρατηρούμε πως το σκορ του αλγορίθμου όσον αναφορά τον Simplicity είναι στο 0.70 σε σχέση με το 1 που είχε ο alpha miner, βέβαια εύκολα κοιτώντας το Petri net μπορούμε να συμπεράνουμε πως είναι λογικό το αποτέλεσμα του heuristic αλγορίθμου δεν είναι το ίδιο απλό με του alpha miner.

8.1.3 Inductive Miner

Τέλος θα κάνουμε την ίδια διαδικασία και για έναν ακόμα αλγόριθμο, τον Inductive miner. Λόγω της επεκτασιμότητας και της ευελιξίας ο αλγόριθμος Inductive miner είναι αυτή τη στιγμή ένας από τους καλύτερους αλγορίθμους για την ανακάλυψη διαδικασιών και χρησιμοποιείται πολύ συχνά για αυτόν το σκοπό. Αυτός ο αλγόριθμος κάνει εκτεταμένη χρήση κρυφών μεταβάσεων, ειδικά για τα μέρη του μοντέλου που υπάρχει κάποιος κύκλος ή κάποιο «άλμα».



Εικόνα 15: Petri net με inductive αλγορίθμου



Εικόνα 16: Process tree του inductive αλγορίθμου

Αποτελέσματα Fitness score του αλγορίθμου Inductive Miner		
perc_fit_traces	**average_trace_fitness**	**log_fitness**
100.0	1.0	1.0

Το συμπέρασμα που βγάζουμε από αυτά τα αποτελέσματα είναι πως το γραφικό μοντέλο που κατασκευάστηκε φαίνεται πως αντικατοπτρίζει πλήρως τα δεδομένα του αρχείου που του δώσαμε ως είσοδο.

Αποτελέσματα Precision, Generalization και Simplicity για τον Inductive Miner		
Precision	**Generalization**	**Simplicity**
0.1137257315258514	0.9591744210019145	0.6277056277056278

Στα συγκεκριμένα αποτελέσματα άξιο αναφοράς είναι πως έχουμε χαμηλότερο precision score σε σχέση με τον heuristic, κάτι όμως που είναι λογικό αν θυμηθούμε τι ακριβώς λαμβάνει υπόψιν το συγκεκριμένο σκορ. Όπως αναφέραμε στο κεφάλαιο 7 το precision σκορ ελέγχει αν το μοντέλο που δημιουργήθηκε απεικονίζει κάποια συμπεριφορά η οποία να δεν υπάρχει στα δεδομένα, επομένως γνωρίζοντας πως ο inductive miner βρίσκει και κρυφές μεταβάσεις είναι λογικό πως το precision score θα είναι χαμηλότερο.

8.2 Μέρος δεύτερο - Εξόρυξη διεργασιών και Bayesian Networks

8.2.1 Δημιουργία του Μπεϋζιανού δικτύου

Όπως αναφέρουμε και πιο πάνω επιλέχθηκε ένα αρχείο καταγραφής δεδομένων που περιέχει πληροφορίες για τραπεζικά δεδομένα και ήταν μέρος του BPI Challenge 2017. Το συγκεκριμένο αρχείο είναι αρκετά μεγάλο σε μέγεθος, κάτι που καθυστερεί την υπόλοιπη διαδικασία. Όμως παρατηρώντας τα δεδομένα που κρατάει το αρχείο βλέπουμε πως μεγάλο μέρος των δεδομένων δεν είναι αναγκαίο για την εργασία μας. Επομένως για να διευκολύνουμε τη διαδικασία φιλτράρουμε το αρχικό αρχείο κρατώντας μόνο της πληροφορίες που είναι σημαντικές για το σκοπό που θέλουμε το αρχείο. Αφού αφαιρέσαμε τις περιττές πληροφορίες και κατ' επέκταση μειώσαμε και τον όγκο του αρχείου πλέον διαχωρίζουμε το αρχικό αρχείο σε δύο αρχεία, ένα για την εκμάθηση του δικτύου και ένα για τις μετέπειτα δοκιμές. Παρακάτω παρουσιάζουμε ένα παράδειγμα event του αρχείου πριν και μετά το φιλτράρισμα.

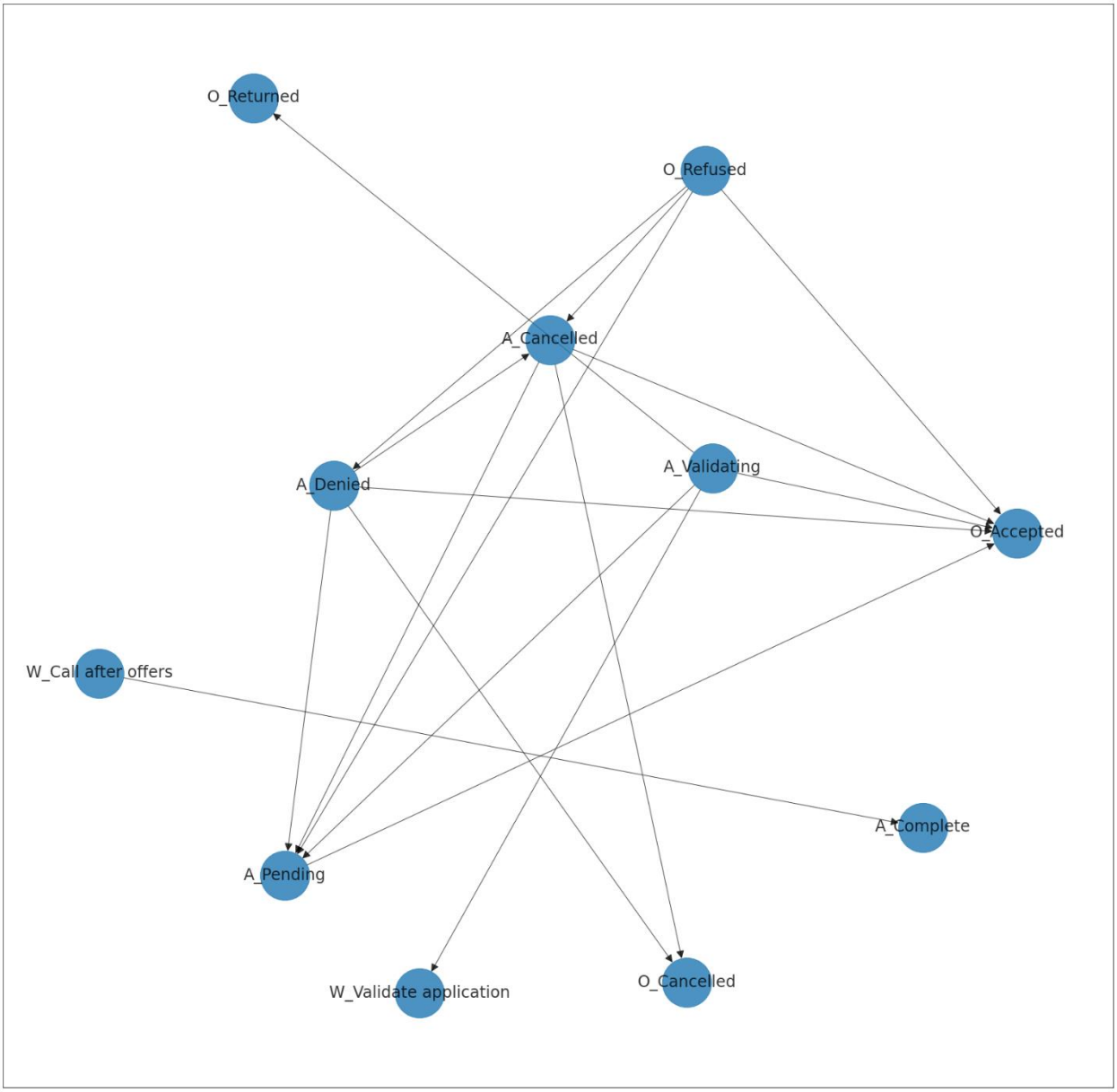
Παράδειγμα event πριν και μετά το φιλτράρισμα	
Event πριν το φιλτράρισμα	{'Action': 'Created', 'org:resource': 'User_1', 'concept:name': 'A_Create Application', 'EventOrigin': 'Application', 'EventID': 'Application_652823628', 'lifecycle:transition': 'complete', 'time:timestamp': datetime.datetime(2016, 1, 1, 9, 51, 15, 304000, tzinfo=datetime.timezone.utc)}
Event μετά το φιλτράρισμα	{'concept:name': 'A_Create Application', 'time:timestamp': datetime.datetime(2016, 1, 1, 9, 51, 15, 304000, tzinfo=datetime.timezone.utc)}

Το αρχείο αποτελείται από ίχνη (traces) που και αυτά με τη σειρά τους αποτελούνται από μια σειρά γεγονότων (events). Στόχος της εργασίας είναι να μπορεί να γίνει πρόβλεψη για την παρουσία ή απουσία ενός event μέσα σε ένα trace. Δηλαδή θα πρέπει να μετατρέψουμε τα δεδομένα μας έτσι ώστε για κάθε ίχνος να ξέρουμε ποια event συνέβησαν και ποια όχι. Ο

πίνακας που ακολουθεί παρουσιάζει για τα 5 πρώτα ίχνη ένα μέρος των δραστηριοτήτων για τα οποία κρατήσαμε τη κατάστασή τους, σκοπός είναι να δείξουμε την αλλαγή από το trace στον πίνακα που θα χρησιμοποιηθεί για την δημιουργία του Bayesian network και όχι η παρουσίαση του συνόλου των δεδομένων.

	A_Accepted	**A_Cancelled**	**A_Complete**	**A_Concept**	**A_Create Application**	**A_Denied**	**A_Incomplete**
0	present	absent	present	present	present	absent	present
1	present	absent	present	present	present	present	absent
2	present	absent	present	present	present	absent	present
3	present	absent	present	present	present	absent	present
4	present	present	present	present	present	absent	absent

Το πρόβλημα που έχουμε αυτή τη στιγμή είναι πως έχουμε τα δεδομένα αλλά δεν γνωρίζουμε τη δομή η οποία τα περιγράφει, για να κατασκευάσουμε ένα Μπεϋζιανό δίκτυο θα πρέπει με κάποιο τρόπο από τα δεδομένα αυτά να φτάσουμε σε ένα γραφικό μοντέλο. Για να πετύχουμε κάτι τέτοιο χρησιμοποιούμε αλγορίθμους μάθησης δομής που μας παρέχει η βιβλιοθήκη pgmpy. Πιο συγκεκριμένα στη περίπτωσή μας χρησιμοποιήσαμε έναν constrain based αλγόριθμο τον PC.



Εικόνα 17 : Αποτέλεσμα εκτέλεσης του αλγορίθμου PC

Όπως βλέπουμε το γραφικό μοντέλο που δημιουργήθηκε δεν περιέχει όλες τις δραστηριότητες που αρχικά είχαμε στο αρχείο, προφανώς γιατί ο αλγόριθμος δεν μπόρεσε να συσχετίσει τις απούσες δραστηριότητες με οποιαδήποτε άλλη. Εφόσον έχουμε το γραφικό μοντέλο πλέον το μόνο στοιχείο που λείπει για να έχουμε έτοιμο το Μπεϋζιανό δίκτυο είναι οι πίνακες πιθανοτήτων για κάθε κόμβο του γραφήματος. Οι πίνακες υπολογίζονται και αυτοί με την βοήθεια της βιβλιοθήκης pgmpy και παρουσιάζονται παρακάτω.

A_Validating(absent)	A_Validating(present)
0.3068	0.6931

O_Refused(absent)	O_Refused(present)
0.8838	0.1161

W_Call after offers(absent)	W_Call after offers(present)
0.0043	0.9956

W_Call after offers	absent	present
A_Complete(absent)	1.0	0.0
A_Complete(present)	0.0	1.0

A_Validating	absent	present
W_Validate application(absent)	1.0	0.0
W_Validate application(present)	0.0	1.0

A_Validating	absent	present
O_Returned(absent)	1.0	0.0049
O_Returned(present)	0.0	0.9950

O_Refused	absent	present
A_Denied(absent)	0.9989	0.0
A_Denied(present)	0.0010	1.0

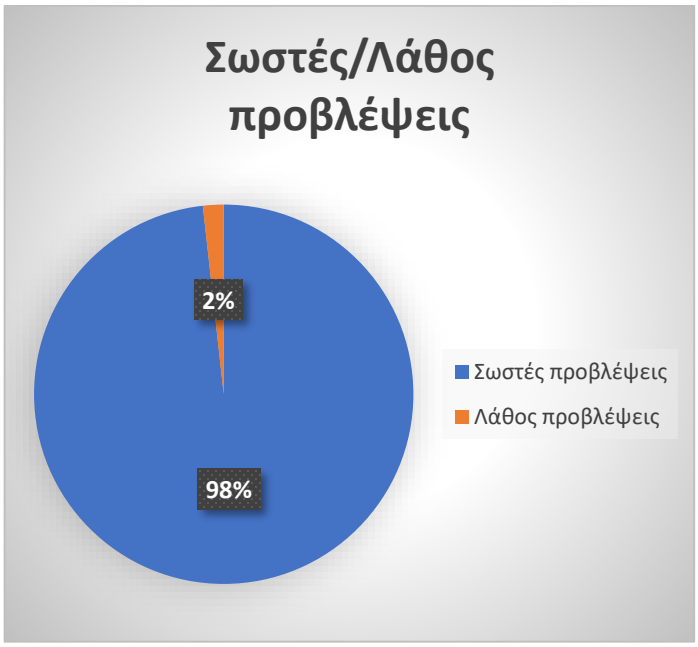
A_Denied	absent	absent	present	present
O_Refused	absent	present	absent	present
A_Cancelled(absent)	0.6211	0.5	1.0	1.0
A_Cancelled(present)	0.3788	0.5	0.0	0.0

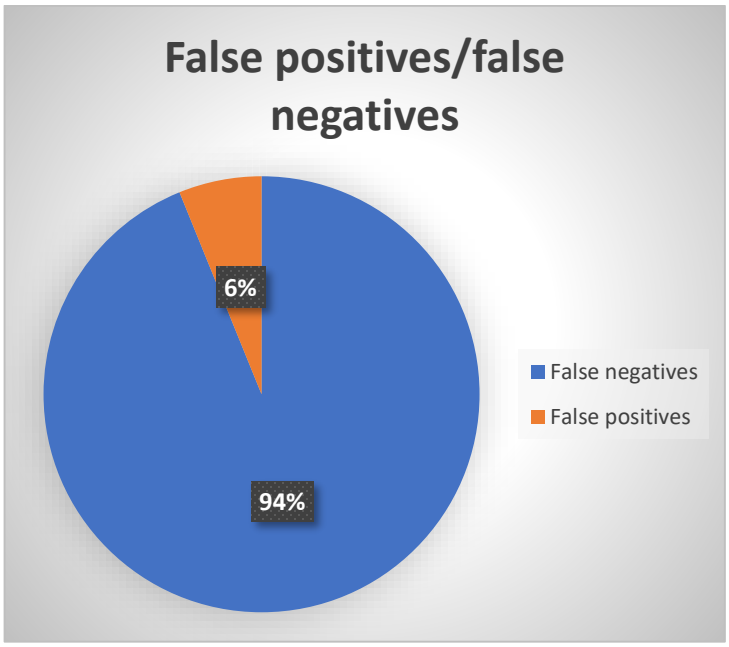
A_Cancelled	absent	absent	present	present
A_Denied	absent	present	absent	present
O_Cancelled(absent)	0.7098	0.9508	0.0	0.5
O_Cancelled(present)	0.2901	0.0491	1.0	0.5

8.2.2 Δοκιμές

Δραστηριότητα	Αριθμός προβλέψεων	Σωστές προβλέψεις	Λάθος προβλέψεις	False negatives	False positives	Ποσοστό επιτυχίας
A_Denied	6301	6250	51	6	45	0.992
A_Cancelled	6301	6295	6	0	6	0.999
O_Accepted	6301	0	0	0	0	1
O_Cancelled	6301	5206	1095	1095	0	0.826
A_Pending	6301	0	0	0	0	1
A_Validating	6301	6256	45	45	0	0.992
W_Validate application	6301	0	0	0	0	1
O_Returned	6301	6285	16	0	16	0.997
W_Call after offers	6301	6301	0	0	0	1
A_Complete	6301	6301	0	0	0	1
O_Refused	6301	6293	8	0	8	0.999
Συνολικά	69311	68090	1221	1146	75	0.982

Τα αποτελέσματα των δοκιμών έχουν ένα εξαιρετικό ποσοστό επιτυχίας 98.2%, ενώ στις λάθος προβλέψεις βλέπουμε πως είναι κυρίως false negatives.





9 Συμπεράσματα και μελλοντική εργασία

Σκοπός της συγκεκριμένης εργασίας είναι να ερευνήσει κατά πόσο μπορούν να χρησιμοποιηθούν τα Μπεϋζιανά δίκτυα για την διαχείριση της αβεβαιότητας στην εξόρυξη διαδικασιών. Αναφερθήκαμε στις βασικές γνώσεις που χρειάζονται για την εύκολη κατανόηση του αντικειμένου της εργασίας. Αρχικά ορίσαμε το βασικό πλαίσιο πάνω στο οποίο εργαστήκαμε, δηλαδή την διαχείριση επιχειρησιακών διαδικασιών και την εξόρυξη διεργασιών, καθώς και όλες τις απαραίτητες θεωρητικές γνώσεις των Μπεϋζιανών δικτύων. Στη συνέχεια χωρίσαμε το πρακτικό μέρος της εργασία σε δύο μέρη. Στο πρώτο είδαμε πως χρησιμοποιούνται οι κλασικοί αλγόριθμοι εξόρυξης διεργασιών για να ανακαλύψουμε διαδικασίες από ένα αρχείο καταγραφής γεγονότων. Στο δεύτερο μέρος ξεκινώντας από το ίδιο αρχείο καταγραφής γεγονότων δημιουργήσαμε ένα Μπεϋζιανό δίκτυο και προσπαθήσαμε να προβλέψουμε την κατάσταση των δραστηριοτήτων με βάση αυτό το δίκτυο. Τέλος παρουσιάσαμε τα αποτελέσματα των δοκιμών.

Κατά τη διαδικασία της μάθησης δομής παρατηρήσαμε πως ο αλγόριθμος PC δημιούργησε ένα γραφικό μοντέλο που δεν περιείχε όλες τις δραστηριότητες του αρχείου καταγραφής γεγονότων, παρά μόνο ένα μέρος αυτών. Αν και κατά τις δοκιμές το ποσοστό επιτυχίας ήταν εξαιρετικό, άξιο έρευνας είναι το γιατί το γραφικό μοντέλο που κατασκευάστηκε δεν είχε όλους τους κόμβους που περιμέναμε. Επιπλέον έρευνα θα μπορούσε να γίνει και στους αλγορίθμους μάθησης δομής, δηλαδή αν υπάρχουν άλλοι αλγόριθμοι με καλύτερα αποτελέσματα στο κομμάτι του γραφικού μοντέλου, χωρίς όμως να χάνουμε την αξιοπιστία που παρατηρήσαμε κατά την διαδικασία των δοκιμών στον PC αλγόριθμο.

Ένα διαφορετικό μονοπάτι για περαιτέρω έρευνα είναι ο τρόπος με τον οποίο μορφοποιήσαμε τα δεδομένα του αρχείου καταγραφής γεγονότων. Δηλαδή να δούμε αν μπορεί να φτιαχτεί ένα Μπεϋζιανό δίκτυο που να προβλέπει όχι μόνο αν ένα σύνολο δραστηριοτήτων θα συμβεί αλλά και με ποια σειρά. Με πιο απλά λόγια θα μπορούσαμε χρησιμοποιώντας την ίδια προσέγγιση και τον ίδιο αλγόριθμο μάθησης δομής να παρατηρήσουμε αν μπορούμε να έχουμε το ίδιο καλά αποτελέσματα για κάποιο πιο σύνθετο ερώτημα.

10 Βιβλιογραφία

[1] Van der Aalst, W. M. (2013). Business process management: a comprehensive survey. International Scholarly Research Notices, 2013.

[2] Daniel, F., Sheng, Q. Z., & Motahari, H. (Eds.). (2019). Business Process Management Workshops: BPM 2018 International Workshops, Sydney, NSW, Australia, September 9-14, 2018, Revised Papers (Vol. 342). Springer

[3] Van Der Aalst, W. M. (2003, September). Business process management demystified: A tutorial on models, systems and standards for workflow management. In Advanced Course on Petri Nets (pp. 1-65). Springer, Berlin, Heidelberg.

[4] Van Der Aalst, W. M. (2003, September). Business process management demystified: A tutorial on models, systems and standards for workflow management. In Advanced Course on Petri Nets (pp. 1-65). Springer, Berlin, Heidelberg.

[5] Van der Aalst, W. M. (2018). Process discovery from event data: Relating models and logs through abstractions. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(3), e1244.

[7] Ross, S. (2014). A first course in probability. Pearson.

[8] Neapolitan, R. E. (2004). *Learning bayesian networks* (Vol. 38). Upper Saddle River, NJ: Pearson Prentice Hall.

[9] Russell, S., & Norvig, P. (2009). Artificial Intelligence: A Modern Approach. edition.

[10] Kalisch, M., & Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research, 8(3)

[11] Margaritis, D. (2003). *Learning Bayesian network model structure from data*. Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.