Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Ηλεκτρονικων Βιομηχανικων Διαταξεων και Συστηματων Αποφασεων

Ανάλυση και Παραγωγή Δεδομένων Κίνησης Νοσοκομειακού Δικτύου με Χρήση Παραγωγικών Μοντέλων Μηχανικής Μάθησης

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΡΑΝΙΑ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Αποφασεων και Διοικησης
Αθήνα, Μάρτιος 2022

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων
Εργαστήριο Συστημάτων Αποφάσεων και Διοίκησης

Ανάλυση και Παραγωγή Δεδομένων Κίνησης Νοσοκομειακού Δικτύου με Χρήση Παραγωγικών Μοντέλων Μηχανικής Μάθησης

Διπλωματικη Εργασια

του

ΔΗΜΗΤΡΙΟΥ ΚΡΑΝΙΑ

Επιβλέπων : Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Μαρτίου 2022.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Δημήτριος Ασκούνης	Ιωάννης Ψαρράς	Χρυσόστομος Δούκας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2022

(Υπογραφή)

.....
Δημήτριος Κρανιάς
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η ανάγκη για την αντιμετώπιση του κινδύνου των κυβερνοεπιθέσεων στον τομέα της υγείας είναι ένα μείζον θέμα και χρήζει άμεσης αντιμετώπισης. Για να πραγματοποιηθεί αυτό υπάρχουν αρκετοί τρόποι, όμως αυτός που αποτέλεσε κίνητρο για την διπλωματική αυτή, ήταν η παρακολούθηση της κίνησης των διαφόρων χρηστών για τον εντοπισμό ασυνήθιστων συμπεριφορών. Για να επιτευχθεί αυτό, το σύνολο από διαθέσιμα πραγματικά νοσοκομειακά δεδομένα δεν επαρκεί και χρειάστηκαν συνθετικά, αληθοφανή δεδομένα για να το συμπληρώσουν.

Αντικείμενο της παρούσας διπλωματικής αποτελεί η ανάλυση των πραγματικών νοσοκομειακών δεδομένων κίνησης και η παραγωγή συνθετικών δεδομένων με παραγωγικά μοντέλα μηχανικής μάθησης. Στο στάδιο της ανάλυσης, μετά την ανωνυμοποίηση των ευαίσθητων πληροφοριών, εξετάστηκαν οι κατανομές κάποιων πεδίων των δεδομένων ώστε να εξαχθούν χρήσιμα συμπεράσματα για αυτές. Στην συνέχεια, έγινε κατηγοριοποίηση ορισμένων χρηστών που χρησιμοποιούν νοσοκομειακές υπηρεσίες και από τους οποίους θα γίνει η παραγωγή δεδομένων στο τελικό βήμα. Έπειτα πραγματοποιήθηκε μια συσταδοποίηση αυτών των χρηστών (για κίνηση προς συγκεκριμένες υπηρεσίες) για τον έλεγχο του πόσο κοντά είναι οι χρήστες ίδιων υπηρεσιών.

Τέλος, εκπαιδεύτηκαν διάφορα γεννητικά μοντέλα μηχανικής μάθησης, μεταξύ των οποίων και μοντέλα βαθιάς μάθησης (Γεννητικά Ανταγωνιστικά Δίκτυα και Variational Autoncoders), σε διάφορα υποσύνολα των δεδομένων. Με αυτά τα εκπαιδευμένα μοντέλα παρήχθησαν τα συνθετικά δεδομένα τα οποία και αξιολογήθηκαν με ορισμένες μεθόδους και μετρικές. Έτσι, προέκυψαν τα πλεονεκτήματα και τα μειονεκτήματα των μοντέλων και οι περιπτώσεις στις οποίες το καθένα θα ήταν χρήσιμο.

Λέξεις Κλειδιά

NetFlow, Συσταδοποίηση, Διερευνητική Ανάλυση Δεδομένων, Παραγωγή Δεδομένων, Γκαουσιανά Μοντέλα Μείξης, Βαθιά Μάθηση, Γεννητικά Ανταγωνιστικά Δίκτυα, Variational Autoencoders

Abstract

The need for dealing with the danger of cyberattacks in healthcare is an important topic that requires immediate solutions. There are many ways to accomplish this, but the one that was the motivating factor for this thesis, was the monitoring of the traffic flow of various users in order to detect unusual behaviour. In order to achieve this, there was the need for synthetic, realistic data to complement existing real medical data.

The subject of this study is the analysis of real medical data of NetFlow type and the generation of synthetic data with generative machine learning models. In the first part of the analysis, after the anonymization of sensitive data, the distributions of various features/columns of the dataset were examined, in order for useful conclusions to be drawn for them. Subsequently, certain users were categorized based on the usage of hospital services, in order to generate the synthetic data from them in the final step of the thesis. Afterward, a clustering of these users took place (based on traffic to certain services) to observe how similar the users of the same services are.

Finally, several generative machine learning models were trained, among them some were deep learning models (Generative Adversarial Networks and Variational Autoencoders), in various subsets of the dataset. The synthetic data were generated with these trained models, which were evaluated with certain methods and metrics. That way, advantages and disadvantages of the models emerged, alongside the cases in which each of them could prove useful.

Keywords

NetFlow, Clustering, Exploratory Data Analysis, Data Generation, Gaussian Mixture Models, Deep Learning, Generative Adversarial Networks, Variational Autoencoders

Ευχαριστίες

Με την εκπόνηση αυτής της διπλωματικής εργασίας, ολοκληρώνεται ένας πολύ σημαντικός κύκλος της ζωής μου, αυτός των προπτυχιακών σπουδών μου στην σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβιο Πολυτεχνείο. Σε αυτή την διαδρομή με στήριξαν αρκετά άτομα, τα οποία επιθυμώ να ευχαριστήσω.

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή κ. Δημήτριο Ασκούνη, για την εμπιστοσύνη που μου έδειξε αναθέτοντας μου αυτή την διπλωματική εργασία. Παράλληλα, χρωστάω ένα μεγάλο ευχαριστώ στον υποψήφιο διδάκτορα Σωτήρη Πελέκη, για την εξαιρετική καθοδήγηση και τις συμβουλές που μου παρείχε καθ΄ όλη την διάρκεια εκπόνησης της διπλωματικής. Ακόμη, δεν θα μπορούσα να παραλείψω τον συνεργάτη και φίλο, Χρήστο Μπέτζελο, με τον οποίο συνεργαστήκαμε άψογα στα πρώτα στάδια της διπλωματικής, πριν ο καθένας εμβαθύνει στο δικό του θέμα.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου, Κωνσταντίνο και Ελένη, για την αμέριστη αγάπη, στήριξη και κατανόηση που έδειξαν και δείχνουν προς εμένα σε κάθε μου βήμα, καθώς και όλες τις σχέσεις που σύναψα κατά την διάρκεια αυτής της διαδρομής, οι οποίες βοήθησαν ώστε να μου μείνει αξέχαστη.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Κίνητρο

Μία από τις σημαντικότερες προκλήσεις στον τομέα της υγείας στην σύγχρονη ψηφιακή εποχή είναι η κυβερνοασφάλεια. Αυτό οφείλεται στο γεγονός ότι τα ιατρικά δεδομένα των ασθενών αποτελούν στόχο κακόβουλων χρηστών, οι οποίοι επιχειρούν με πολλές μεθόδους κυβερνοεπιθέσεων να τα αποκτήσουν (π.χ. για να τα πουλήσουν σε τρίτους όπως ασφαλιστικές εταιρίες) ή να τα κρυπτογραφήσουν/κλειδώσουν ζητώντας χρήματα για την αποκρυπτογράφηση/ξεκλείδωμα τους (Ransomware attacks). Συνήθως για να επιτύχουν τον σκοπό τους χρησιμοποιούν μεθόδους εξαπάτησης (Social Engineering) με αποδέκτες το υγειονομικό προσωπικό (όπως το Phishing). Στο διάγραμμα 1.1 φαίνονται κάποια στατιστικά από τις αναφορές παραβίασης δεδομένων της Verizon (Verizon Data Breach Investigation Report) για τα έτη 2020 ([1]) και 2021 ([2]) που επιβεβαιώνουν τα παραπάνω. Ο εσωτερικός παράγοντας σαν κίνδυνος και τις δύο χρονιές είναι πολύ υψηλός (48% και 39%) και οι κακόβουλοι χρήστες στοχεύουν κυρίως προσωπικά και ιατρικά δεδομένα για χρηματικά κίνητρα.

Υπάρχουν αρκετοί τρόποι για να αποφευχθεί ή να αντιμετωπιστεί αυτός ο κίνδυνος, μεταξύ των οποίων είναι η χρήση state-of-the-art συστημάτων προστασίας των υπολογιστών και του ιατρικού εξοπλισμού, η κατάλληλη εκπαίδευση του ιατρικού προσωπικού ώστε να μην υποπίπτουν σε τέτοια λάθη και η παρακολούθηση της δικτυακής κίνησης του προσωπικού για να εντοπιστούν τυχούσες περίεργες ¨συμπεριφορές¨, δηλαδή που δεν αντιστοιχούν σε υπό κανονικές συνθήκες δικτυακή συμπεριφορά. Όσον αφορά τον τελευταίο τρόπο αντιμετώπισης όμως, υπάρχουν αρκετά προβλήματα, από την έλλειψη επαρκών δεδομένων για την προσομοίωση φυσιολογικής/μη φυσιολογικής συμπεριφοράς ενός χρήστη μέχρι και την ανάγκη για διατήρηση της ανωνυμιάς του χρήστη και άλλων δεδομένων που θα μπορούσαν να την θέσουν σε κίνδυνο (όπως διευθύνσεις IP, αναζητήσεις στο διαδίκτυο κ.α.). Το πρώτο πρόβλημα προκύπτει άμεσα από το δεύτερο (δεν μπορούν να αντληθούν αρκετά δεδομένα λόγω του **Γενικού Κανονισμού Προστασίας Προσωπικών Δεδομένων** ή αλλιώς **GDPR**), οπότε παρουσιάζεται στην συνέχεια ο σκοπός της διπλωματικής και πώς επιλύθηκαν αυτά τα προβλήματα.

Frequency	798 incidents, 521 with confirmed data disclosure
Top Patterns	Miscellaneous Errors, Web Applications and Everything Else represent 72% of breaches
Threat Actors	External (51%), Internal (48%), Partner (2%), Multiple (1%) (breaches)
Actor Motives	Financial (88%), Fun (4%), Convenience (3%) (breaches)
Data Compromised	Personal (77%), Medical (67%), Other (18%), Credentials (18%) (breaches)

(α′) Στατιστικά για το 2020

Frequency	655 incidents, 472 with confirmed data disclosure
Top Patterns	Miscellaneous Errors, Basic Web Application Attacks and System Intrusion represent 86% of breaches
Threat Actors	External (61%), Internal (39%) (breaches)
Actor Motives	Financial (91%), Fun (5%), Espionage (4%), Grudge (1%) (breaches)
Data Compromised	Personal (66%), Medical (55%), Credentials (32%), Other (20%), (breaches)

(β′) Στατιστικά για το 2021

Σχήμα 1.1: Στατιστικά παραβίασης δεδομένων στον τομέα της υγείας για τα έτη 2020 και 2021 από την Verizon

1.2 Σκοπός Εργασίας

Ο σκοπός αυτής της διπλωματικής εργασίας, σε πρώτο στάδιο, είναι η ανωνυμοποίηση, η επεξεργασία και η ανάλυση των δεδομένων κίνησης νοσοκομειακού δικτύου που αντλήθηκαν σε νοσοκομειακό περιβάλλον από προσωπικό του **Εργαστηρίου Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων** και του νοσοκομείου. Με βάση τα συμπεράσματα που έχουν ληφθεί από το προηγούμενο στάδιο, επιχειρείται η παραγωγή συνθετικών δεδομένων και η αξιολόγηση τους για να μπορέσουν να χρησιμοποιηθούν σε εφαρμογές προσομοίωσης συμπεριφοράς χρηστών.

1.3 Δομή Εργασίας

Η παρούσα διπλωματική εργασία χωρίζεται σε 6 κεφάλαια:

1. **Εισαγωγή:** Στο Κεφάλαιο 1, αναφέρεται το κίνητρο που οδήγησε σε αυτή την διπλωματική και τον σκοπό αυτής, καθώς και μια σύντομη παρουσίαση της δομής της.

2. **Θεωρητικό Υπόβαθρο:** Στο Κεφάλαιο 2, παρουσιάζεται το θεωρητικό υπόβαθρο των τεχνικών που χρησιμοποιήθηκαν στα στάδια της ανάλυσης, παραγωγής και αξιολόγησης των δεδομένων, αλλά και σύντομη επεξήγηση των πρωτοκόλλων που διέπουν τα δεδομένα (όπως NetFlow, DHCP).

3. **Διερευνητική Ανάλυση Δεδομένων:** Στο Κεφάλαιο 3, αρχικά περιγράφεται το σύνολο δεδομένων πριν και μετά τις πρώτες τροποποιήσεις του (που πραγματοποιούνται για την ασφάλεια του νοσοκομειακού προσωπικού). Στην συνέχεια, αναλύονται αυτά τα δεδομένα για την εξαγωγή πληροφοριών και συμπερασμάτων, που βοηθούν στο να γίνει

δυνατή επιπλέον προεπεξεργασία που θα χρειαστεί στα επόμενα βήματα της διπλωματικής εργασίας και ιδίως στο βήμα της παραγωγής δεδομένων.

4. **Συσταδοποίηση:** Στο Κεφάλαιο 4, επιχειρείται μια ομαδοποίηση των νοσοκομειακών χρηστών με διάφορους αλγορίθμους συσταδοποίησης, ώστε να αξιολογηθεί η ποιότητα των δεδομένων μας.

5. **Παραγωγή Δεδομένων:** Στο Κεφάλαιο 5, πραγματοποιείται παραγωγή δεδομένων κίνησης με χρήση παραγωγικών μοντέλων μηχανικής μάθησης και αξιολόγηση αυτών των μοντέλων μέσω της σύγκρισης των παραγόμενων δεδομένων με τα πραγματικά.

6. **Συμπεράσματα-Μελλοντικές Επεκτάσεις:** Τέλος, στο Κεφάλαιο 6 παρουσιάζονται συνοπτικά τα αποτελέσματα και τα συμπεράσματα που προέκυψαν τόσο από την επεξεργασία και ανάλυση του συνόλου δεδομένων καθώς και από την παραγωγή νέων δεδομένων, σε συνδυασμό με πιθανούς τρόπους επέκτασης της παρούσας μελέτης.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Ροή Δικτύου (Network/Traffic Flow)

Όπως αναφέρθηκε και στην εισαγωγή, η διπλωματική αυτή αρχικά εστιάζει στην παρακολούθηση του νοσοκομειακού δικτύου, για την συλλογή των κατάλληλων δεδομένων που θα αναλυθούν και θα αποτελέσουν βάση για την παραγωγή νέων, συνθετικών δεδομένων. Η μορφή αυτών των δεδομένων είναι στην μορφή ροής δικτύου.

Η ροή δικτύου (ή ροή κίνησης ή ροή πακέτων όπως αλλιώς ονομάζεται) είναι μια αλληλουχία πακέτων που μεταφέρει πληροφορίες μεταξύ δύο συσκευών. Βέβαια αυτή είναι αρκετά αφελής εξήγηση, γιατί στην περίπτωση που υπάρχουν περισσότερες από μια είδους επικοινωνίες μεταξύ των συσκευών, τότε δεν θα ξεχώριζαν. Γι΄ αυτό τον λόγο εκτός από τις πληροφορίες που μεταφέρονται από την μια συσκευή στην άλλη, μεταφέρεται και η επικεφαλίδα του πακέτου (packet header), η οποία περιέχει τις κατάλληλες πληροφορίες αυτής της σύνδεσης/μεταφοράς (π.χ. τις IP πηγής και προορισμού, τον αριθμό πρωτοκόλλου, τα ports πηγής και προορισμού κ.α.).

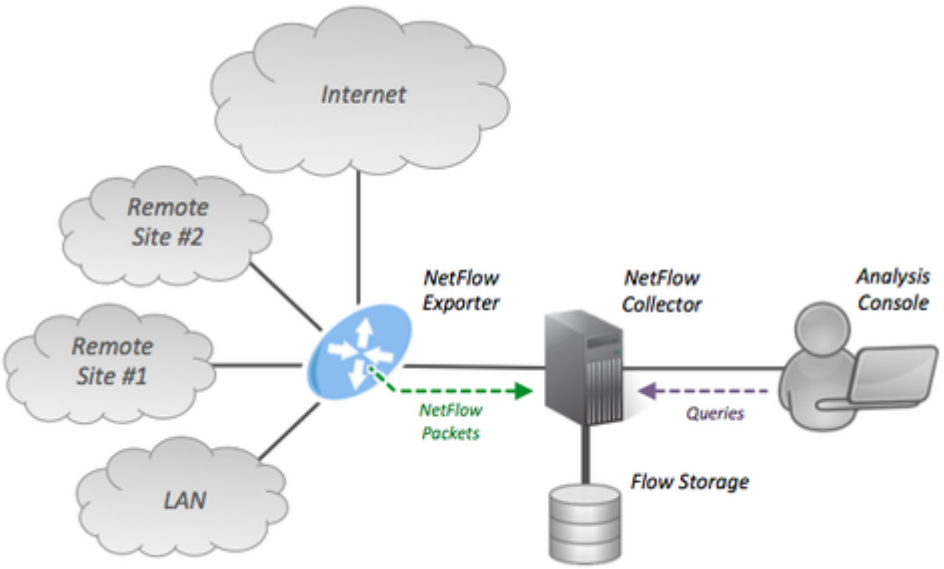
2.1.1 NetFlow

Τα παραπάνω πεδία που χρειάζονται για την μεταφορά της πληροφορίας οδηγούν στην ανάγκη ενός πρωτοκόλλου που θα καθορίζει την επικεφαλίδα του πακέτου (με τα πεδία που επιλέγονται). Για την παρακολούθηση της ροής δικτύου χρησιμοποιήθηκε στην παρούσα διπλωματική το πρωτόκολλο NetFlow.

Το NetFlow είναι ένα σύστημα πρωτοκόλλου δικτύου που αναπτύχθηκε από την Cisco το 1996 για την συλλογή δεδομένων κίνησης δικτύου. Με την ανάλυση αυτών των δεδομένων μπορούν να εξαχθούν πολλά χρήσιμα συμπεράσματα, όπως θα δειχθεί και στην διπλωματική αυτή. Η παρακολούθηση ροής δικτύου συνήθως αποτελείται από τρία συστατικά:

- **Εξαγωγέας ροής:** συγκεντρώνει τα πακέτα σε ροή και τα εξάγει προς έναν ή περισσότερους συλλέκτες ροής.
- **Συλλέκτης ροής:** είναι υπεύθυνος για την υποδοχή, αποθήκευση και προεπεξεργασία των δεδομένων ροής που παρέλαβε από έναν εξαγωγέα ροής.
- **Εφαρμογή ανάλυσης:** αναλύει τα δεδομένα ροής που παραλήφθησαν, όπως προγραμματιστεί από τον χειριστή του.

Τα στάδια αυτά φαίνονται και στο διάγραμμα 2.1.



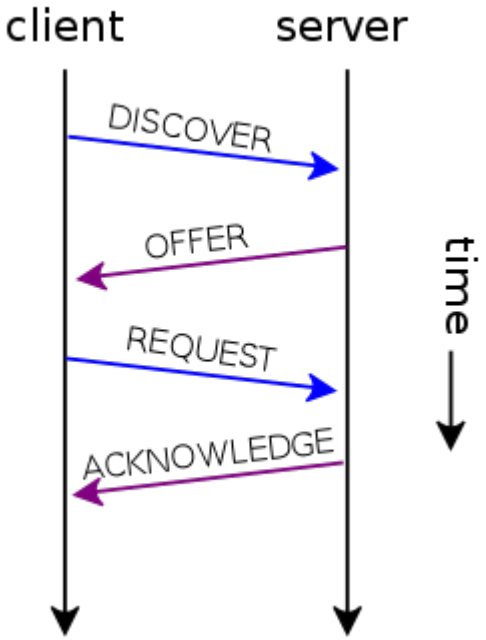
Σχήμα 2.1: Η αρχιτεκτονική του NetFlow

Η έκδοση του πρωτοκόλλου NetFlow που χρησιμοποιήθηκε είναι η 9η (έναντι της εξίσου δημοφιλούς 5ης). Αυτή η επιλογή έγινε διότι τα πεδία που μεταφέρονται με την ροή στην περίπτωση της 9ης έκδοσης είναι αρκετά περισσότερα και μπορούν να επιλεγούν δυναμικά, έναντι στα στατικά πεδία που έχει η 5η έκδοση (δηλαδή στην 5η έκδοση τα πεδία είναι προκαθορισμένα). Έτσι υπήρξε η επιλογή για πεδία που περιείχαν αρκετές χρήσιμες πληροφορίες, που στην περίπτωση της 5ης έκδοσης δεν θα υπήρχε.

2.2 Dynamic Host Configuration Protocol

Το Dynamic Host Configuration Protocol ή εν συντομία DHCP είναι ένα πρωτόκολλο διαχείρισης δικτύου που χρησιμοποιείται σε IP δίκτυα για την αυτόματη ή δυναμική ανάθεση διευθύνσεων IP και άλλων παραμέτρων επικοινωνίας σε συσκευές που είναι συνδεδεμένες στο δίκτυο και χρησιμοποιούν αρχιτεκτονική client-server.

Αναλυτικότερα, όταν μια συσκευή συνδέεται στο δίκτυο, το λογισμικό του DHCP Client εκπέμπει ένα ερώτημα (query) όπου ζητάει τις απαραίτητες πληροφορίες. Οποιοσδήποτε DHCP Server μπορεί να εξυπηρετήσει αυτή την αίτηση και να απαντήσει, δίνοντας τις πληροφορίες που έχει διαμορφώσει ο διαχειριστής (όπως IP Address, Domain Name, Default Getaway κ.α.) για κάποιο συγκεκριμένο χρονικό διάστημα. Με το πέρας αυτού του χρονικού διαστήματος, ο DHCP Client ζητάει αρχικά τις ίδιες παραμέτρους (lease renewal), αλλά ο DHCP Server μπορεί να επιλέξει διαφορετικές, ανάλογα με την πολιτική ανάθεσης που έχουν ορίσει οι διαχειριστές. Στο σχήμα 2.2 που ακολουθεί, φαίνεται μια συνηθισμένη λειτουργία του πρωτοκόλλου DHCP που μόλις περιγράψαμε.



Σχήμα 2.2: Απεικόνιση λειτουργίας του πρωτοκόλλου DHCP

2.3 Αλγόριθμοι Συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης, οι οποίοι αποτελούν υποκατηγορία της μη επιβλεπόμενης μάθησης, αφορούν το πρόβλημα του διαχωρισμού ενός συνόλου δεδομένων σε ξεχωριστές ομάδες, τις συστάδες. Η συστάδα είναι μια περιοχή στον χώρο δεδομένων, της οποίας τα σημεία θεωρείται ότι βρίσκονται κοντά μεταξύ τους (δηλαδή ότι έχουν κοινά χαρακτηριστικά), συγκριτικά με σημεία άλλων συστάδων. Βέβαια, το ποια σημεία θα κατηγοριοποιηθούν μαζί και το πόσο κοντά θα είναι εξαρτάται από την επιλογή του αλγορίθμου συσταδοποίησης και από τις παραμέτρους που θα του δώσουμε. Εξίσου σημαντική είναι και η κατάλληλη προεπεξεργασία των δεδομένων τόσο για την βέλτιστη εκτέλεση των αλγορίθμων, όσο και για την οπτικοποίηση των αποτελεσμάτων σε 2 ή 3 διαστάσεις, με τεχνικές όπως η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis) και η κανονικοποίηση των τιμών των δεδομένων (Normalization ή Standardization).

Οι γενικές κατηγορίες αλγορίθμων συσταδοποίησης είναι:

- **Αλγόριθμοι βασισμένοι στα κεντροϊδή**, στους οποίους κάθε συστάδα ορίζεται από ένα κεντροϊδές σημείο (το οποίο δεν χρειάζεται να ανήκει στα δεδομένα μας). Τα κεντροϊδή ορίζονται με τέτοιον τρόπο ώστε να ελαχιστοποιείται η απόσταση των σημείων από το κεντροϊδές. Ο πιο γνωστός αλγόριθμος αυτής της κατηγορίας είναι ο αλγόριθμος k-Means.

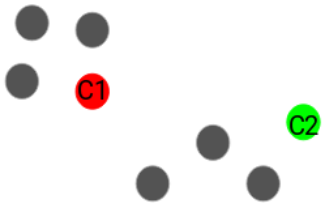
- **Αλγόριθμοι βασισμένοι σε κατανομές**, στους οποίους γίνεται η υπόθεση ότι τα δεδομένα ανήκουν σε κατανομές (όπως για παράδειγμα οι Γκαουσιανές κατανομές). Οι συγκεκριμένοι αλγόριθμοι είναι πιθανοτικοί, δηλαδή κάθε κατανομή αναθέτει μια πιθανότητα σε ένα σημείο (του πόσο πιθανό είναι να ανήκει σε καθεμία από αυτές) και το σημείο αυτό τοποθετείται στην κατανομή με την μεγαλύτερη πιθανότητα.

- **Αλγόριθμοι βασισμένοι στην πυκνότητα**, στους οποίους οι συστάδες ορίζονται σαν περιοχές με υψηλή πυκνότητα δεδομένων και σε αντίθεση με άλλους αλγορίθμους, τα σημεία σε αραιές περιοχές θεωρούνται ακραίες τιμές (outliers) και δεν αποδίδονται σε κάποια συστάδα. Χαρακτηριστικό παράδειγμα τέτοιου αλγορίθμου είναι ο αλγόριθμος DBSCAN.
- **Αλγόριθμοι βασισμένοι στην συνεκτικότητα** ή αλλιώς **ιεραρχικοί αλγόριθμοι**, οι οποίοι επιχειρούν να χτίσουν μια ιεραρχία από συστάδες, είτε συσσωρευτικά (agglomerative clustering) είτε διαιρετικά (divisive clustering).

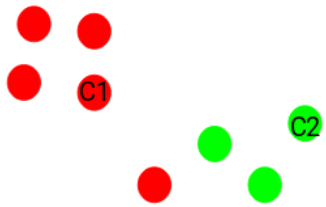
2.3.1 Αλγόριθμος k-Means

Ο αλγόριθμος k-Means πρόκειται για έναν από τους πιο διαδεδομένους αλγορίθμους συσταδοποίησης, χάρη στις πολλαπλές υλοποιήσεις που διατίθενται στο διαδίκτυο και στην ευκολία στην χρήση του (αρκεί να οριστεί μόνο ο αριθμός k των συστάδων). Τα βήματα του αλγορίθμου k-Means είναι τα εξής:

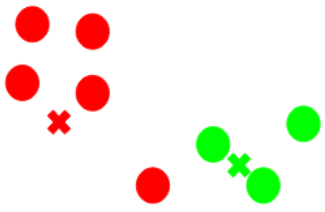
1. Επιλέγονται k τυχαία σημεία σαν κεντροϊδή σημεία των συστάδων. 2.3α΄
2. Κάθε σημείο των δεδομένων ανατίθεται στην συστάδα του κοντινότερου κεντροϊδούς σημείου. 2.3β΄
3. Για κάθε συστάδα υπολογίζεται το νέο κεντροϊδές σημείο, του οποίου η θέση καθορίζεται από την μέση τιμή των σημείων της συστάδας. 2.3γ΄
4. Τα βήματα 2, 3 επαναλαμβάνονται μέχρι να επιτευχθεί σύγκλιση του αλγορίθμου (δηλαδή να μην αλλάξει συστάδα κανένα σημείο σε μια επανάληψη). 2.3δ΄



(α΄) Επιλογή k κεντροϊδών



(β΄) Ανάθεση κάθε σημείου στο κοντινότερο κεντροϊδές



(γ΄) Υπολογισμός των νέων κεντροϊδών

(δ΄) Επανάληψη των βημάτων 2, 3 μέχρι την σύγκλιση

Σχήμα 2.3: Τα βήματα του αλγορίθμου k-Means

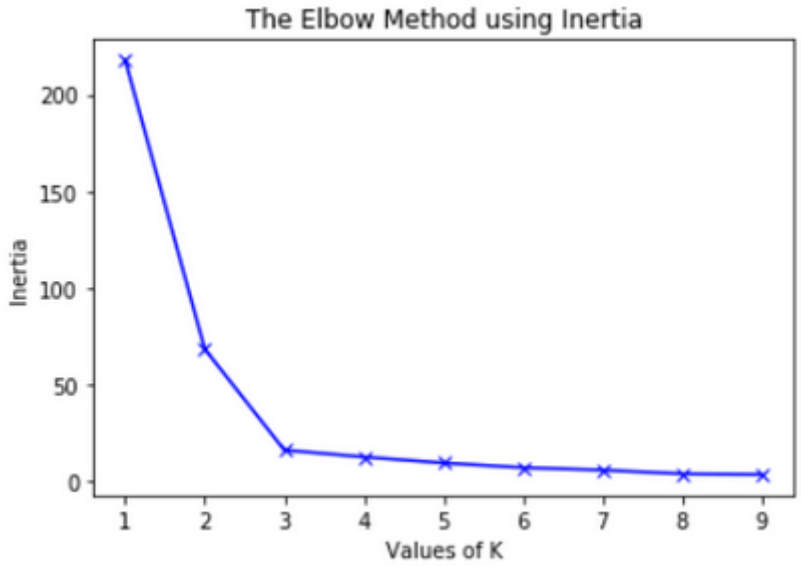
Επειδή ο αριθμός των συστάδων καθορίζεται πριν την εκτέλεση του αλγορίθμου, δεν είναι σίγουρο αν τα δεδομένα θα διαχωριστούν βέλτιστα. Για να αντιμετωπιστεί αυτό το πρόβλημα στην παρούσα διπλωματική, χρησιμοποιείται η ΄μέθοδος του αγκώνα΄ (Elbow Method). Η μέθοδος αυτή έχει ως σκοπό να βρει τον ιδανικό αριθμό συστάδων εκτελώντας τον αλγόριθμο για ένα προεπιλεγμένο εύρος τιμών και υπολογίζοντας κάθε φορά την αδράνεια (inertia) ή όπως αλλιώς αποκαλείται, το άθροισμα των τετραγωνικών αποστάσεων των σημείων από το κεντροϊδές τους (within-cluster sum of squared distances).

$$\sum_{i=1}^{k} \sum_{\mu_j \in C_i} (||x_i - \mu_j||^2)$$

όπου,

- k ο αριθμός των συστάδων που έχει προεπιλεγεί
- x_i το κεντροϊδές της συστάδας i
- μ_j το σημείο j μιας συστάδας
- C_i το σύνολο σημείων της συστάδας i

Το k που εν τέλει θα επιλεγεί δεν είναι αυτό για το οποίο ελαχιστοποιείται το παραπάνω άθροισμα, αλλά αυτό για το οποίο περαιτέρω αύξηση του θα φέρει μικρή μείωση στο άθροισμα. Αυτή η επιλογή μπορεί να γίνει εμπειρικά, βλέποντας για ποιο k η μείωση γίνεται γραμμική από εκθετική (όπως φαίνεται και για k = 3 στο διάγραμμα 2.4 παρακάτω). Μπορεί να γίνει όμως και αυτόματα χωρίς την επίβλεψη κάποιου χρήστη, με υλοποιήσεις όπως αυτή που περιγράφεται στο [3], όπου η επιλογή του βέλτιστου k γίνεται με βάση την κυρτότητα της καμπύλης.



Σχήμα 2.4: Παράδειγμα χρήσης του Elbow Method

2.3.2 Μοντέλα Μείξης (Mixture Models)

Τα μοντέλα μείξης είναι πιθανοτικά μοντέλα, τα οποία έχουν σαν στόχο την ομαδοποίηση των δεδομένων σε συστάδες, όχι όμως με βάση κάποια συνάρτηση απόστασης (όπως με τους άλλους αλγορίθμους συσταδοποίησης που παρουσιάζονται στην διπλωματική), αλλά με βάση την πιθανότητα ενός σημείου να ανήκει σε κάποια κατανομή της μείξης (και ανατίθεται σε αυτή με την μεγαλύτερη).

Για την σύσταση ενός Γενικού Μοντέλου Μείξης χρειάζονται συνήθως τα παρακάτω:

- N τυχαίες μεταβλητές που έχουν παρατηρηθεί, κατανεμημένες σε μια μείξη K συστατικών (components), με τα components να ανήκουν στην ίδια οικογένεια κατανομών (π.χ. όλες Γκαουσιανές, όλες Μπερνούλι κ.ο.κ.), αλλά με διαφορετικές παραμέτρους.
- N τυχαίες κρυφές μεταβλητές που προσδιορίζουν την ταυτότητα του component μείξης κάθε σημείου, όπου η καθεμία είναι κατανεμημένη με βάση μια $K-$διάστατη κατηγορική κατανομή.
- K βάρη μείξης, που είναι πιθανότητες που αθροίζουν στο 1
- K παράμετροι, όπου η κάθεμια θα προσδιορίζει τις παραμέτρους του αντίστοιχου component μείξης.

Πιο αναλυτικά, ένα τυπικό μοντέλο μείξης έχει τις παραμέτρους:

K	=	αριθμός των παραμέτρων μείξης
N	=	αριθμός των παρατηρήσεων/σημείων δεδομένων
$\theta_{i=1...K}$	=	παράμετροι της κατανομής ενός σημείου που σχετίζονται με το component i
$\phi_{i=1...K}$	=	βάρος μείξης του component i
$\boldsymbol{\phi}$	=	K-διάστατος πίνακας όλων των $\phi_{1...K}$· πρέπει να αθροίζει στο 1
$z_{i=1...N}$	=	component της παρατήρησης i
$x_{i=1...N}$	=	παρατήρηση i
$F(x\|\theta)$	=	κατανομή πιθανότητας ενός σημείου, παραμετροποιημένη ως προς το θ
$z_{i=1...N}$	$\sim$	Categorical($\boldsymbol{\phi}$)
$x_{i=1...N}\|z_{i=1...N}$	$\sim$	$F(\theta_{z_i})$

Υπάρχουν αρκετές κατανομές (F) οι οποίες μπορούν να χρησιμοποιηθούν για την δημιουργία ενός μοντέλου μείξης. Η πιο συνηθισμένη από αυτές είναι η Γκαουσιανή ή Normal (την οποία και χρησιμοποιούμε στην παρούσα διπλωματική) και μερικές άλλες γνωστές είναι η Μπερνούλι, η Δυωνυμική και η Log-Normal.

Ένα τυπικό Γκαουσιανό μοντέλο μείξης περιγράφεται από τα παρακάτω:

K, N	=	όπως παραπάνω
$\phi_{i=1...K}, \boldsymbol{\phi}$	=	όπως παραπάνω
$z_{i=1...N}, x_{i=1...N}$	=	όπως παραπάνω
$\theta_{i=1...K}$	=	$\{\mu_{i=1...K}, \sigma^2_{i=1...K}\}$
$\mu_{i=1...K}$	=	μέσος όρος της παραμέτρου μείξης i
$\sigma^2_{i=1...K}$	=	διακύμανση της παραμέτρου μείξης i
$z_{i=1...N}$	$\sim$	Categorical($\boldsymbol{\phi}$)
$x_{i=1...N}$	$\sim$	$\mathcal{N}(\mu_{z_i}, \sigma^2_{z_i})$

Το παραπάνω όμως Γκαουσιανό μοντέλο μείξης δεν μπορεί να επεξεργαστεί δεδομένα περισσότερων από 1 διαστάσεων. Για τον λόγο αυτό, χρησιμοποιούνται τα πολυμεταβλητά Γκαουσιανά μοντέλα μείξης (Multivariate Gaussian Mixture Models), που είναι ικανά να απεικονίσουν κατανομές μεγαλύτερων διαστάσεων. Η μόνη διαφορά στην σημειογραφία είναι ότι η κατανομή των πολυμεταβλητών Γκαουσιανών είναι $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}$, όπου $\boldsymbol{\mu})$ ο $k-$διάστατος πίνακας μέσων τιμών και $\boldsymbol{\Sigma}$ ο $k \times k$ πίνακας συνδιακυμάνσης, έναντι του $\mathcal{N}(\mu, \sigma^2)$ των απλών Γκαουσιανών μοντέλων, όπου μ, σ^2 είναι οι μονοδιάστατοι πίνακες μέσης τιμής και διακύμανσης.

Οι νέες παράμετροι των μοντέλων μείξης υπολογίζονται με την βοήθεια μιας αρκετά δημοφιλούς μεθόδου, του αλγορίθμου Expectation-Maximization (EM Algorithm). Ο αλγόριθμος EM αποτελείται από 2 βήματα, το E-Step και το M-Step, αφού πρώτα γίνει τυχαία αρχικοποίηση των παραμέτρων. Στο βήμα Expectation, ο αλγόριθμος προσπαθεί να μαντέψει την τιμή του $z^{(i)}$ βασιζόμενος στις παραμέτρους, ενώ στο βήμα Maximization ενημερώνει τις τιμές των παραμέτρων βασιζόμενος στην μαντεψιά του $z^{(i)}$. Αυτά τα 2 βήματα επαναλαμβάνονται μέχρι την σύγκλιση. Τα βήματα του αλγορίθμου EM για τα Γκαουσιανά μοντέλα μείξης, εκφρασμένα με μαθηματικές συναρτήσεις, παρουσιάζονται παρακάτω.

E-Step:

- $w_j^{(i)} := p\left(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma\right)$

M-Step:

- $\phi_j := \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}$
- $\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$
- $\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} w_j^{(i)}}$

Με τον κανόνα του Bayes προκύπτουν για το E-Step:

- $p\left(x^{(i)} | z^{(i)} = j; \mu, \Sigma\right) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \mu_j\right)^T \Sigma_j^{-1} \left(x^{(i)} - \mu_j\right)\right)$
- $p\left(z^{(i)} = j; \phi\right) = \phi_j$

Όπως και στην περίπτωση του αλγορίθμου k-Means, για την επιλογή του βέλτιστου αριθμού παραμέτρων μείξης (που πρέπει να καθοριστεί πριν την εκτέλεση του αλγορίθμου), επιλέγουμε δύο κριτήρια πληροφορίας, το Bayesian Information Criterion (BIC) και το Akaike Information Criterion (AIC). Τα κριτήρια αυτά ενώ μοιάζουν αρκετά, έχουν μια σημαντική διαφορά που θα εξηγηθεί παρακάτω. Παρ΄ όλα αυτά είναι κατάλληλα για την επιλογή ενός μοντέλου, το οποίο θεωρούν καλύτερο, μεταξύ πεπερασμένου πλήθους μοντέλων (όπου στην περίπτωση αυτή, διαφορετικό μοντέλο σημαίνει διαφορετικός αριθμός παραμέτρων μείξης). Ο σκοπός τους είναι να εισάγουν μια ποινή στην αύξηση των παραμέτρων, διότι όταν εκπαιδεύονται μοντέλα με περισσότερες παραμέτρους, αυξάνεται η πιθανοφάνεια (likelihood), αλλά αυξάνεται και η πιθανότητα υπερπροσαρμογης (overfitting) των μοντέλων αυτών.

Το κριτήριο BIC:

$$\text{BIC} = k\ln(n) - 2\ln(\widehat{L})$$

όπου,

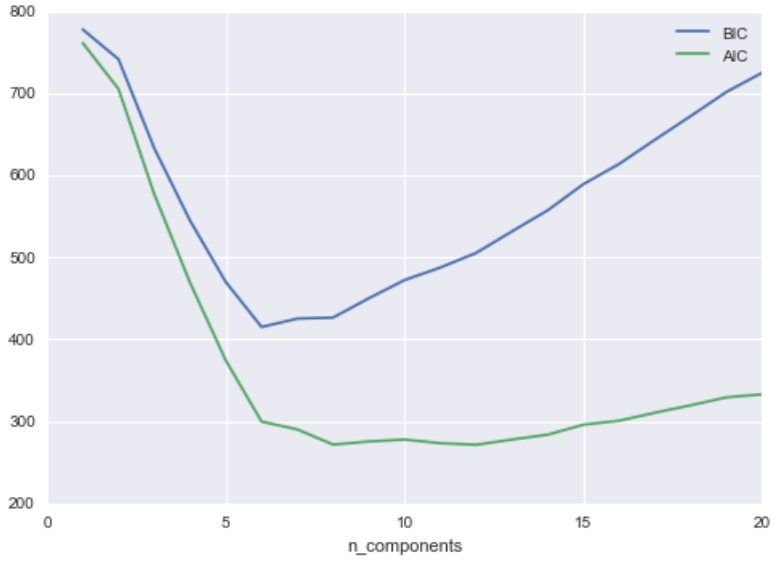
- $\hat{L}$ = η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας του μοντέλου. Για παράδειγμα, $\hat{L} = p(x \mid \widehat{\theta}, M)$, όπου $\widehat{\theta}$ οι τιμές των παραμέτρων που μεγιστοποιούν την συνάρτηση πιθανοφάνειας και M το μοντέλο
- n = ο αριθμός των σημείων του συνόλου δεδομένων
- k = ο αριθμός των παραμέτρων που εκτιμώνται από το μοντέλο

Το κριτήριο AIC:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

όπου,

- $\hat{L}$ = όπως παραπάνω
- k = όπως παραπάνω



Σχήμα 2.5: Παράδειγμα γραφικών παραστάσεων των κριτηρίων πληροφορίας BIC, AIC

Η σημαντικότερη διαφορά, που φαίνεται και στο διάγραμμα 2.5, είναι ότι η ποινή που εισάγει το AIC είναι αρκετά πιο επιεικής από αυτή του BIC, καθώς δεν εξαρτάται από τον αριθμό των δειγμάτων n. Αυτό μπορεί πολλές φορές να οδηγήσει σε overfitting (κάτι που θα έπρεπε να αντιμετωπίσει αυτή η μέθοδος), ενώ με το BIC στο ακριβώς αντίθετο (underfitting). Μια άλλη διαφορά είναι ότι το AIC προσπαθεί να βρει το μοντέλο που περιγράφει καλύτερα μια άγνωστη, πολυδιάστατη πραγματικότητα, η οποία δεν βρίσκεται ποτέ στα μοντέλα που εξετάζονται, ενώ το BIC θέλει να βρει το καλύτερο μοντέλο το οποίο θεωρεί ότι είναι ένα από τα μοντέλα που εξετάζεται.

2.3.3 Ιεραρχική Συσσωρευτική Συσταδοποίηση (Hierarchical Agglomerative Clustering)

Η ιεραρχική συσσωρευτική συσταδοποίηση, η οποία αποτελεί την πιο συνηθισμένη μέθοδο ιεραρχικής συσταδοποίησης (έναντι της ιεραρχικής διαιρετικής συσταδοποίησης), είναι μια "bottom-up" προσέγγιση, δηλαδή ξεκινάει με αριθμό συστάδων όσες και το πλήθος των δεδομένων και σταδιακά ενώνονται τα δύο κοντινότερα clusters σε κάθε επανάληψη, έως ότου όλα ενωθούν σε ένα cluster.

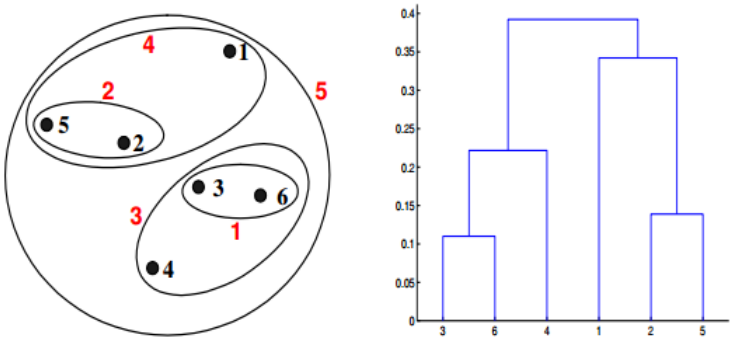
Η απόφαση του αλγορίθμου σχετικά με το ποιες δύο συστάδες θα ενωθούν σε κάθε βήμα, εξαρτάται από δύο παράγοντες, την σύνδεση (linkage) και την μετρική απόστασης που θα χρησιμοποιήσει η σύνδεση.

Συνηθισμένες μετρικές απόστασης είναι η Ευκλείδια απόσταση ($\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$), η τετραγωνική Ευκλείδια ($\|a-b\|_2 = \sum_i (a_i - b_i)^2$), η Manhattan ($\|a-b\|_1 = \sum_i |a_i - b_i|$) και η Malahanobis ($\sqrt{(a-b)^\top S^{-1}(a-b)}$, όπου S ο πίνακας συνδιακύμανσης). Η μετρική όμως που θα χρησιμοποιηθεί στην ανάλυση δεν είναι καμία από τις παραπάνω και είναι η Wasserstein Distance, η οποία θα αναλυθεί στην επόμενη ενότητα του παρόντος κεφαλαίου και ο λόγος που χρησιμοποιήθηκε αυτή στο κεφάλαιο 4).

Οι πιο διαδεδομένοι τύποι συνδέσεων είναι:

- Η πλήρης σύνδεση (complete linkage), υπολογίζει την απόσταση μεταξύ των πιο απομακρυσμένων σημείων των clusters και ενώνει αυτά με την μικρότερη.
- Η απλή σύνδεση (single linkage), υπολογίζει την απόσταση μεταξύ των πιο κοντινών σημείων των clusters και ενώνει αυτά με την μικρότερη.
- Η σύνδεση μέσου όρου (average linkage), υπολογίζει το άθροισμα των αποστάσεων όλων των πιθανών δυάδων σημείων διά τον αριθμό των δυάδων και ενώνει αυτά με την μικρότερη.
- Άλλοι γνωστοί τύποι συνδέσεων είναι η σύνδεση κεντροϊδών (centroid linkage), το κριτήριο του Ward, η σύνδεση μέσου όρου με βάρη κ.α.

Για την απεικόνιση της ιεραρχίας των clusters και την επιλογή του βέλτιστου αριθμού clusters, χρησιμοποιούνται δενδρογράμματα (όπως φαίνονται στο διάγραμμα 2.6 παρακάτω) ώστε να διακρίνονται και οι αποστάσεις μεταξύ των ιεραρχικών επιπέδων.



Σχήμα 2.6: Παράδειγμα ιεραρχικής συσσωρευτικής συσταδοποίησης με το δενδρόγραμμά της

2.4 Wasserstein Distance

Η απόσταση Wasserstein ή μετρική Kantorovich-Rubinstein είναι μια συνάρτηση απόστασης ορισμένη μεταξύ κατανομών πιθανοτήτων σε ένα δοσμένο μετρικό χώρο M. Πήρε το όνομα της από τον Ρώσο μαθηματικό Leonid Vaserstein από την έρευνα του [4] το 1969, αλλά η μετρική είχε ήδη οριστεί από τον Ρώσο μαθηματικό-οικονομολόγο Leonid Kantorovich, το 1939 στην μελέτη του πάνω στην βέλτιστη μετακίνηση (optimal transport) αγαθών και υλικών [5].

Ο ορισμός της p^{th} Wasserstein Distance μεταξύ δύο κατανομών πιθανοτήτων a, b μπορεί να δωθεί ως εξής:

$$W_p(a,b) = (\min_{\gamma \in \mathbb{R}_+^{m \times n}} \sum_{i,j} \gamma_{i,j} \|x_i - y_j\|_p)^{\frac{1}{p}}$$

όπου,

- $\gamma_{i,j}$ ο πίνακας βέλτιστης μετακίνησης
- $\sum_i \gamma_{i,j} = b_j;\ \sum_j \gamma_{i,j} = a_i;\ \gamma_{i,j} \geq 0$
- $x_i \in a,\ y_j \in b$

Ένας τρόπος για να γίνει κατανοητός ο παραπάνω ορισμός είναι να ληφθεί υπόψη το πρόβλημα βέλτιστης μετακίνησης. Μια κατανομή με μάζα $\mu(x)$, σε ένα χώρο $\mathbb{X}$, θέλουμε να μετακινηθεί με τέτοιον τρόπο ώστε να μετατραπεί στην κατανομή $\nu(y)$ στον ίδιο χώρο. Για να γίνει αυτό, θεωρούμε χωρίς βλάβη της γενικότητας, ότι οι κατανομές μ, ν έχουν συνολική μάζα ίση με 1. Επίσης θεωρούμε μια συνάρτηση κόστους $c(x,y) \mapsto [0,\infty)$ που δίνει το κόστος για την μεταφορά μονάδας μάζας από το σημείο x στο σημείο y. Το σχέδιο μεταφοράς της μ στην ν περιγράφεται από την συνάρτηση $\gamma(x,y)$. Για να έχει νόημα αυτό το σχέδιο, πρέπει να ισχύουν:

- $\int \gamma(x,y)\,\mathrm{d}y = \mu(x)$
- $\int \gamma(x,y)\,\mathrm{d}x = \nu(y)$

Το σχέδιο γ όμως δεν είναι μοναδικό. Το βέλτιστο σχέδιο μεταφοράς είναι το σχέδιο με το μικρότερο κόστος από όλα τα πιθανά σχέδια μεταφοράς. Οπότε το κόστος του βέλτιστου σχεδίου μεταφοράς είναι:

$$C = \inf_{\gamma \in \Gamma(\mu,\nu)} \int c(x,y)\,\mathrm{d}\gamma(x,y)$$

Αν το κόστος για μια μετακίνηση είναι απλά η απόσταση μεταξύ των δύο σημείων, τότε το βέλτιστο κόστος ταυτίζεται με την W_1 απόσταση.

2.5 Γεννητικά Μοντέλα (Generative Models)

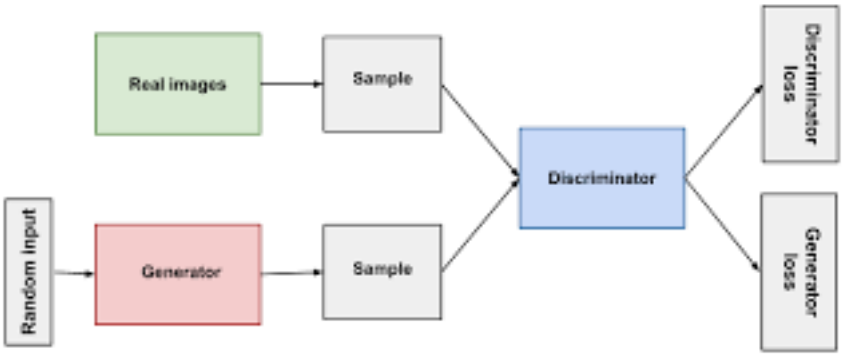
Η γεννητική μοντελοποίηση (generative modeling) είναι μια διαδικασία μη επιβλεπόμενης μάθησης, η οποία περιλαμβάνει την εκμάθηση μοτίβων (patterns) στα δεδομένα εισόδου, τέτοια ώστε να μπορεί να χρησιμοποιηθεί για την παραγωγή νέων δεδομένων που θα προκύπτουν από τα δεδομένα εισόδου. Χαρακτηριστικά γεννητικά μοντέλα είναι τα προαναφερθέντα Γκαουσιανά μοντέλα μείξης, τα κρυφά μοντέλα Markov, τα γεννητικά ανταγωνιστικά δίκτυα, οι variational autoencoders κ.α.

2.5.1 Γεννητικά Ανταγωνιστικά Δίκτυα (GANs)

Τα Γεννητικά Ανταγωνιστικά Δίκτυα ή GANs (Generative Adversarial Networks), που πρωτοεμφανίστηκαν στην δημοσίευση [6] του Goodfellow, είναι μοντέλα βαθιάς μηχανικής μάθησης, στα οποία δύο νευρωνικά δίκτυα, ο γεννήτορας (generator) και ο διευκρινιστής (discriminator), αντιμετωπίζουν ο ένας τον άλλο σε ένα ΄παιχνίδι΄ όπου το κέρδος του ενός είναι η απώλεια του άλλου (zero-sum game).

Ο γεννήτορας έχει σαν σκοπό την παραγωγή αληθοφανών δεδομένων, ενώ ο διευκρινιστής να μάθει να διαχωρίζει τα πραγματικά δεδομένα από τα ψεύτικα που παράγει ο γεννήτορας.

Η αρχιτεκτονική του GAN παρουσιάζεται στο διάγραμμα 2.7.



Σχήμα 2.7: Παράδειγμα αρχιτεκτονικής GAN

Όπως φαίνεται και στο διάγραμμα, ο διευκρινιστής συνδέεται τόσο με την συνάρτηση απωλειών του, όσο και με αυτή του γεννήτορα. Όταν εκπαιδεύεται ο διευκρινιστής, αγνοείται αυτή του γεννήτορα και το αντίθετο συμβαίνει όταν εκπαιδεύεται ο γεννήτορας.

Κατά την διάρκεια της εκπαίδευσης του διευκρινιστή, συμβαίνουν τα εξής βήματα:

1. Ο διευκρινιστής ταξινομεί τα ψεύτικα δεδομένα του γεννήτορα, αλλά και τα αληθινά.
2. Η συνάρτηση απωλειών τον τιμωρεί για κάθε λάθος ταξινόμηση (δηλαδή αν ταξινομήσει ένα αληθινό δεδομένο για ψεύτικο και το αντίστροφο).
3. Τα βάρη του διευκρινιστή ενημερώνονται μέσω backpropagation από την συνάρτηση απωλειών μέσω του δικτύου του διευκρινιστή.

Κατά την διάρκεια της εκπαίδευσης του γεννήτορα:

1. Αρχικά γίνεται παραγωγή τυχαίου Γκαουσιανού θορύβου.
2. Στην συνέχεια γίνεται επεξεργασία του θορύβου από τον γεννήτορα και παραγωγή αποτελέσματος.
3. Μετά ο διευκρινιστής αποφαίνεται αν το παραγόμενο αποτέλεσμα είναι αληθινό ή όχι.
4. Αν αποφανθεί ότι είναι ψεύτικο τότε η συνάρτηση απωλειών τιμωρεί τον γεννήτορα.
5. Τέλος, αποκτούνται τα gradients μέσω backpropagation από το δίκτυο του γεννήτορα και του διευκρινιστή και ενημέρωση των βαρών του γεννήτορα με αυτά.

Να τονιστεί ότι κατά την εκπαίδευση του κάθε νευρωνικού (γεννήτορα ή διευκρινιστή) για 1 ή περισσότερα βήματα (π.χ. για ένα βήμα εκπαίδευσης του γεννήτορα γίνονται 5 βήματα εκπαίδευσης του διευκρινιστή), το άλλο νευρωνικό παραμένει σταθερό. Η εκπαίδευση και των 2 επαναλαμβάνεται για τον αριθμό των εποχών που έχει οριστεί και ιδανικά θέλουμε να υπάρχει σύγκλιση (δηλαδή ο διευκρινιστής θα βάζει label στην τύχη). Πολλές φορές, υπάρχει ο κίνδυνος, αν συνεχιστεί η εκπαίδευση μετά την σύγκλιση, ο γεννήτορας να χειροτερέψει (επειδή εκπαιδεύεται σε διευκρινιστή που ταξινομεί στην τύχη).

Ο γεννήτορας και ο διευκρινιστής εκπαιδεύονται και αξιολογούνται με συναρτήσεις απωλειών μεταξύ πραγματικών και ψεύτικων δεδομένων. Στη αρχική δημοσίευση του Goodfellow προτάθηκε η minimax συνάρτηση απωλειών που φαίνεται παρακάτω:

$$\min_G \max_D V(D,G) = \mathop{\mathbb{E}}_{x \sim p_{data}(x)}[\log(D(x))] + \mathop{\mathbb{E}}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

όπου,

- $D(x)$ είναι η εκτίμηση της πιθανότητας του διευκρινιστή ότι ένα αληθινό δεδομένο είναι αληθινό
- $D(G(z))$ είναι η εκτίμηση της πιθανότητας του διευκρινιστή ότι ένα ψεύτικο δεδομένο είναι αληθινό
- Η συνάρτηση απωλειών του διευκρινιστή: $\log(D(x)) + \log(1 - D(G(z)))$
- Η συνάρτηση απωλειών του γεννήτορα: $\log(1 - D(G(z)))$

Μια άλλη αρκετά διαδεδομένη συνάρτηση απωλειών είναι η απώλεια Wasserstein, και τα GANs που την υλοποιούν ονομάζονται Wasserstein GANs ή WGANs. Αυτή η συνάρτηση απωλειών αναφέρθηκε για πρώτη φορά στην δημοσίευση [7] και στην συνέχεια βελτιώθηκε με την προσθήκη gradient penalty στην δημοσίευση [8].

Η διαφοροποίηση της από την minimax είναι ότι σε αυτή την περίπτωση ο διευκρινιστής δεν ταξινομεί τα στιγμιότυπα, γιατί εξάγει έναν αριθμό, ο οποίος δεν είναι ανάμεσα στο 0 και στο 1 ώστε να τεθεί για κατώφλι το 0.5 που θα καθορίζει το αν ένα στιγμιότυπο είναι πραγματικό ή ψεύτικο. Απλά προσπαθεί να κάνει τον αριθμό αυτό μεγαλύτερο για τα πραγματικά δεδομένα απ΄ ότι για τα ψεύτικα. Για τον παραπάνω λόγο, ο διευκρινιστής στην περίπτωση του WGAN ονομάζεται κριτικός (critic), διότι δεν διευκρινίζει στην πραγματικότητα μεταξύ αληθινού και ψεύτικου.

Η συνάρτηση αυτή έχει την μορφή:

$$\min_G \max_{D \in \mathcal{D}} V(D, G) = \mathop{\mathbb{E}}_{x \sim p_{data}(x)}[D(x)] - \mathop{\mathbb{E}}_{z \sim p_z(z)}[D(G(z)]$$

όπου,

- $\mathcal{D}$ είναι το σύνολο των 1−Lipschitz συναρτήσεων για τις οποίες ισχύει ο δυισμός Kantorovich-Rubinstein
- Η συνάρτηση απωλειών του κριτικού: $D(x) - D(G(z))$
- Η συνάρτηση απωλειών του γεννήτορα: $D(G(z))$

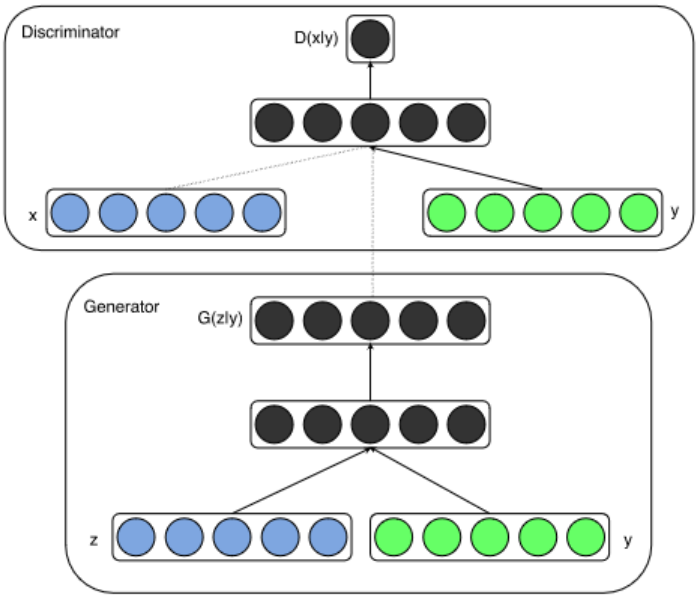
Με αυτή την συνάρτηση απωλειών βελτιώνεται αρκετά η ευστάθεια του GAN και αποφεύγονται προβλήματα όπως το mode collapse και το vanishing gradient. Επίσης όπως αναφέρθηκε στην προηγούμενη ενότητα, η απόσταση Wasserstein αποτελεί μετρική, σε αντίθεση με το cross-entropy loss που χρησιμοποιείται συνήθως στην minimax.

2.5.1.1 Conditional GANs

Τα Conditional GANs αποτελούν επέκταση των απλών GANs για τις περιπτώσεις όπου υπάρχει κάποια κλάση ή επισήμανση (label) y. Αυτή η πληροφορία ενοποιείται με τον θόρυβο εισόδου σε μια κοινή αναπαράσταση που στην συνέχεια δίνεται στον γεννήτορα. Η συνάρτηση απωλειών minimax με αυτή την προσέγγιση γίνεται:

$$\min_G \max_D V(D, G) = \mathop{\mathbb{E}}_{x \sim p_{data}(x)}[\log(D(x \mid y))] + \mathop{\mathbb{E}}_{z \sim p_z(z)}[\log(1 - D(G(z \mid y)))]$$

Μια περιγραφή της αρχιτεκτονικής του CGAN από την δημοσίευση [9] παρουσιάζεται στο διάγραμμα 2.8.



Σχήμα 2.8: Παράδειγμα αρχιτεκτονικής CGAN

2.5.1.2 CTGAN

Τα GANs έχουν αρκετά μεγάλη επιτυχία στην παραγωγή αληθοφανών εικονών και γι΄ αυτό χρησιμοποιούνται και σε άλλα πεδία, όπως και στα δεδομένα τύπου NetFlow που εξετάζονται στην παρούσα διπλωματική. Επειδή όμως τα δεδομένα μορφής πίνακα (tabular data) έχουν αρκετές ιδιαιτερότητες, η ανάγκη για κατάλληλη προεπεξεργασία των δεδομένων είναι τεράστια. Μερικά από τα προβλήματα που αντιμετωπίζουν τα GANs είναι τα:

- Τα δεδομένα αποτελούνται από διαφορετικούς τύπους δεδομένων (όπως αριθμητικά, κατηγορικά, boolean κ.α. Για αυτό τον λόγο χρειάζεται ειδική προεπεξεργασία.
- Στις εικόνες, τα pixel ακολουθούν κατανομές που μοιάζουν με Γκαουσιανές, οπότε μπορούν να κανονικοποιηθούν στο $[-1, 1]$. Οι αριθμητικές τιμές στα tabular data δεν μοιράζονται αυτή την ιδιότητα, οπότε μπορεί να οδηγήσουν σε vanishing gradient προβλήματα.
- Αρκετά ανισόρροπα κατηγορικά δεδομένα, δηλαδή μερικές κλάσεις στα δεδομένα είναι αρκετά πιο συχνές από τις άλλες και μπορεί να οδηγήσει σε mode collapse, γιατί ο διευκρινιστής δεν θα μπορεί να αναγνωρίσει κλάσεις με λίγες εμφανίσεις και ο γεννήτορας μπορεί να τον ξεγελάει συνέχεια.

Για να αντιμετωπιστούν τα παραπάνω προβλήματα στη δημοσίευση [10] αναπτύχθηκε η αρχιτεκτονική CTGAN (Conditional Tabular Generative Adversarial Network).

Αρχικά πραγματοποιήθηκε η κατάλληλη προεπεξεργασία των δεδομένων, που ονομάστηκε mode-specific normalization. Σε αυτή την προεπεξεργασία τα κατηγορικά δεδομένα απεικονίζονται ως one-hot-encoded πίνακες όπως συνηθίζεται ($\mathbf{d}_{i,j}$), αλλά όσον αφορά τα συνεχή δεδομένα δεν πραγματοποιήθηκε η συνηθισμένη min-max κανονικοποίηση μεταξύ $[-1, 1]$. Στην θέση της χρησιμοποιήθηκε η εξής μέθοδος:

- Αρχικά σε κάθε συνεχή στήλη δεδομένων εκπαιδεύτηκε μια variational Γκαουσιανή κατανομή (VGM) και από αυτή προέκυψαν τα modes, έστω η_k με βάρη μ_k και τυπική απόκλιση ϕ_k.
- Για κάθε τιμή $c_{i,j}$ της στήλης C_i υπολογίζεται η πιθανότητα να προέρχεται από το κάθε mode.
- Έτσι δειγματοληπτόντας από το πιθανό mode για κάθε τιμή και κανονικοποιόντας σύμφωνα με τα χαρακτηριστικά του προκύπτουν τα $\alpha_{i,j} = \frac{c_{i,j}-\eta_k}{\phi_k}$, $\beta_{i,j} = [0, \ldots, 1, \ldots, 0]$ όπου 1 παίρνει στην θέση k που υποδεικνύεται από το mode.

Έτσι για κάθε γραμμή στα δεδομένα προκύπτει η αναπαράσταση:

$$r_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \ldots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus \mathbf{d}_{1,j} \oplus \ldots \oplus \mathbf{d}_{N_d,j}$$

Η αρχιτεκτονική του Conditional Generator φαίνεται στο διάγραμμα 2.9.
όπου,

- cond είναι ο πίνακας που προκύπτει από όλους τους one-hot-encoded πίνακες των κατηγορικών στηλών, ενωμένων σειριακά.
- FC είναι τα πλήρως συνδεδεμένα layers
- ReLU, gumbel, tanh είναι συναρτήσεις ενεργοποίησης
- BN είναι η κανονικοποίηση ενός batch

$$
\begin{aligned}
h_0 &= z \oplus cond \\
h_1 &= h_0 \oplus \texttt{ReLU}(\texttt{BN}(\texttt{FC}_{|cond|+|z|\to 256}(h_0))) \\
h_2 &= h_1 \oplus \texttt{ReLU}(\texttt{BN}(\texttt{FC}_{|cond|+|z|+256\to 256}(h_1))) \\
\hat{\alpha}_i &= \texttt{tanh}(\texttt{FC}_{|cond|+|z|+512\to 1}(h_2)) && 1 \le i \le N_c \\
\hat{\beta}_i &= \texttt{gumbel}_{0.2}(\texttt{FC}_{|cond|+|z|+512\to m_i}(h_2)) && 1 \le i \le N_c \\
\hat{\mathbf{d}}_i &= \texttt{gumbel}_{0.2}(\texttt{FC}_{|cond|+|z|+512\to |D_i|}(h_2)) && 1 \le i \le N_d
\end{aligned}
$$

Σχήμα 2.9: Αρχιτεκτονική του γεννήτορα του CTGAN

$$
\begin{aligned}
h_0 &= \mathbf{r}_1 \oplus \ldots \oplus \mathbf{r}_{10} \oplus cond_1 \oplus \ldots \oplus cond_{10} \\
h_1 &= \texttt{drop}(\texttt{leaky}_{0.2}(\texttt{FC}_{10|\mathbf{r}|+10|cond|\to 256}(h_0))) \\
h_2 &= \texttt{drop}(\texttt{leaky}_{0.2}(\texttt{FC}_{256\to 256}(h_1))) \\
\mathcal{C}(\cdot) &= \texttt{FC}_{256\to 1}(h_2)
\end{aligned}
$$

Σχήμα 2.10: Αρχιτεκτονική του διευκρινιστή του CTGAN

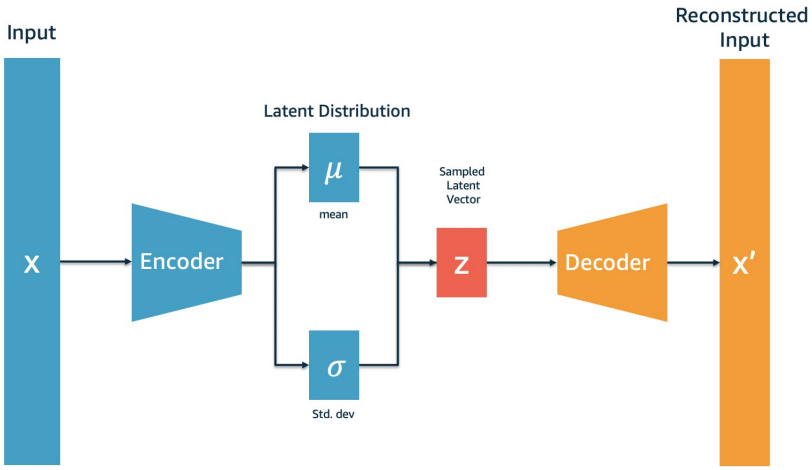
Η αρχιτεκτονική του Discriminator φαίνεται στο διάγραμμα 2.10 και χρησιμοποιεί την αρχιτεκτονική PacGAN [11] με 10 δείγματα για να αποφευχθεί το mode collapse.

όπου,

- Drop είναι το Drop layer
- leaky είναι η Leaky ReLU συνάρτηση ενεργοποίησης

2.5.2 Variational Autoencoders (VAE)

Οι Variational Autoencoders αποτελούν μια μέθοδο βαθιάς μάθησης, η οποία προτάθηκε από τους Kingma, Welling στο [12], που μοιάζει σε αρχιτεκτονική με τους Autoencoders, αλλά τόσο ο στόχος τους όσο και η υλοποίηση τους διαφέρει σημαντικά. Οι Variational Autoencoders σκοπεύουν στην συμπίεση των δεδομένων εισόδου σε μια κρυμμένη πολυμεταβλητή κατανομή (το κομμάτι του encoding), ώστε να ανακατασκευαστούν με όσο το δυνατόν μεγαλύτερη ακρίβεια (το κομμάτι του decoding). Η αρχιτεκτονική του φαίνεται στο διάγραμμα 2.11:



Σχήμα 2.11: Αρχιτεκτονική Variational Autoencoder

Θεωρώντας δεδομένα εισόδου $\mathbf{x}$ που χαρακτηρίζονται από μια άγνωστη κατανομή $p(\boldsymbol{\xi})$, ο στόχος είναι να μοντελοποιηθεί αυτή η κατανομή μέσω μιας παραμετροποιημένης κατανομής p_θ. Αν θεωρήσουμε σαν $\mathbf{z}$ τον πίνακα του κρυφού επιπέδου (latent space), τότε η κατανομή γίνεται:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z})\, d\mathbf{z}$$

όπου $p_\theta(\mathbf{x}, \mathbf{z})$ είναι η από κοινού κατανομή των $\mathbf{x}$, $\mathbf{z}$. Με τον κανόνα της αλυσίδας προκύπτει:

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p_\theta(\mathbf{x} \mid \mathbf{z}) p_\theta(\mathbf{z})\, d\mathbf{z}$$

Έτσι μπορούν να οριστούν οι σχέσεις μεταξύ των δεδομένων εισόδου και της κρυφής αναπαράστασης ως εξής:

- Εκ των προτέρων πιθανότητα $p_\theta(\mathbf{z})$
- Πιθανοφάνεια $p_\theta(\mathbf{x}|\mathbf{z})$
- Εκ των υστέρων πιθανότητα $p_\theta(\mathbf{z}|\mathbf{x})$

Για τον υπολογισμό της εκ των υστέρων πιθανότητας χρησιμοποιείται ο κανόνας του Bayes. Δυστυχώς όμως η πιθανότητα $p_\theta(x)$ που χρειάζεται για τον υπολογισμό της, είναι από υπολογιστικά πολύ δύσκολο να βρεθεί έως και αδύνατο. Για αυτόν τον λόγο επιχειρείται η προσέγγιση του $p_\theta(\mathbf{z}|\mathbf{x})$ με την συνάρτηση $q_\phi(\mathbf{z}|\mathbf{x})$.

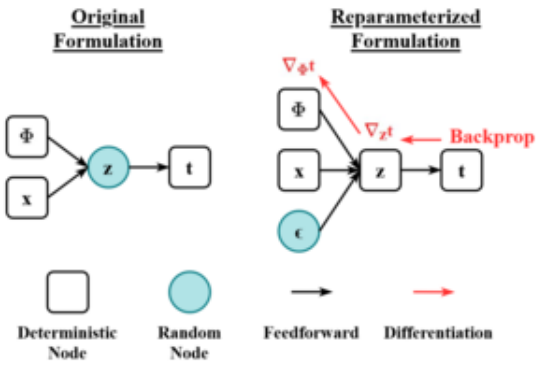
Για να βρεθεί το $q_\phi(\mathbf{z} \mid \mathbf{x})$ (που είναι $\approx q_\phi(\mathbf{z})$) χρησιμοποιείται το Evidence Lower Bound (ή εν συντομία ELBO), το οποίο είναι ένα κάτω φράγμα για την κατανομή $p_\theta(\mathbf{x})$ και γράφεται ως εξής:

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z}\sim q_\phi}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})}\right]$$

Το ELBO πρόκειται για το δεξί μέλος της παραπάνω ανίσωσης. Υπολογίζεται μαθηματικά χρησιμοποιώντας την ανίσωση Kullback-Leibler. Ο μαθηματικός τύπος για την συνάρτηση απωλειών ELBO δίνεται από την παρακάτω σχέση:

$$L_{\theta,\Phi} = -\log(p_\theta(\mathbf{x})) + D_{KL}(q_\Phi(\mathbf{z}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})) =$$
$$= -E_{\mathbf{z}\sim q_\Phi(\mathbf{z}|\mathbf{x})}(\log(p_\theta(\mathbf{x} \mid \mathbf{z}))) + D_{KL}(q_\Phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z}))$$

Τέλος, για να γίνει ο υπολογισμός του ELBO κατάλληλος για την εκπαίδευση του VAE, χρησιμοποιείται το Reparameterization Trick. Αυτό χρειάζεται για να γίνουν εφικτές backpropagation διεργασίες, όπως ο στοχαστικός αλγόριθμος καθοδικής κλίσης (stochastic gradient descent). Αρχικά γίνεται η θεώρηση ότι το κρυφό επίπεδο ακολουθεί Γκαουσιανή κατανομή $\mathbf{z} \sim q_\Phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Για αυτή την περίπτωση ένα κατάλληλο reparameterization είναι το $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ τέτοιο ώστε $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}$. Έτσι, όπως φαίνεται και στο διάγραμμα 2.12, μπορεί να γίνει backpropagate η κλίση (gradient) χωρίς να εμπλακεί η στοχαστική μεταβλητή κατά την ενημέρωση.



Σχήμα 2.12: Η αλλαγή στην σύνθεση μετά το Reparameterization Trick

2.5.2.1 TVAE

Εκτός της αρχιτεκτονικής CTGAN που προτείνεται στο [10], προτείνεται και μια αρχιτεκτονική για Variational Autoencoders στο ίδιο paper, που έχει σκοπό να αντιμετωπίσει τα ίδια προβλήματα που αντιμετώπιζε το απλό GAN στα tabular data. Αυτή η μέθοδος ονομάζεται Tabular Variational Autoencoder ή TVAE εν συντομία. Για τα δεδομένα εισόδου πραγματοποιείται η ίδια προεπεξεργασία (mode-specific normalization). Μια γραμμή δεδομένων είναι όπως προαναφέρθηκε και στο CTGAN:

$$r_j = \alpha_{1,j} \oplus \beta_{1,j} \oplus \ldots \oplus \alpha_{N_c,j} \oplus \beta_{N_c,j} \oplus \mathbf{d}_{1,j} \oplus \ldots \oplus \mathbf{d}_{N_d,j}$$

Η αρχιτεκτονική του $p_\theta(r_j|z_j)$ για αυτό τον λόγο χρειαζόταν αλλαγή σε σχέση με τον απλό VAE. Έτσι παρουσιάζεται στο διάγραμμα 2.13 η νέα αρχιτεκτονική που παράγει $2 * N_c + N_d$ μεταβλητές. Θεωρείται ότι το $\alpha_{i,j}$ ακολουθεί Γκαουσιανή κατανομή με διαφορετικό μέσο όρο και απόκλιση, ενώ τα $\beta_{i,j}$, $\mathbf{d}_{i,j}$ ακολουθούν κατηγορική κατανομή.

$$\begin{aligned}
h_1 &= \texttt{ReLU}(\texttt{FC}_{128\to128}(z_j)) \\
h_2 &= \texttt{ReLU}(\texttt{FC}_{128\to128}(h_1)) \\
\bar{\alpha}_{i,j} &= \texttt{tanh}(\texttt{FC}_{128\to1}(h_2)) && 1 \le i \le N_c \\
\hat{\alpha}_{i,j} &\sim \mathcal{N}(\bar{\alpha}_{i,j}, \delta_i) && 1 \le i \le N_c \\
\hat{\beta}_{i,j} &\sim \texttt{softmax}(\texttt{FC}_{128\to m_i}(h_2)) && 1 \le i \le N_c \\
\hat{\mathbf{d}}_{i,j} &\sim \texttt{softmax}(\texttt{FC}_{128\to|D_i|}(h_2)) && 1 \le i \le N_d \\
p_\theta(\mathbf{r}_j|z_j) &= \textstyle\prod_{i=1}^{N_c} \mathbb{P}(\hat{\alpha}_{i,j} = \alpha_{i,j}) \prod_{i=1}^{N_c} \mathbb{P}(\hat{\beta}_{i,j} = \beta_{i,j}) \prod_{i=1}^{N_d} \mathbb{P}(\hat{\alpha}_{i,j} = \alpha_{i,j})
\end{aligned}$$

Σχήμα 2.13: Αρχιτεκτονική του αποκωδικοποιητή του TVAE

Η αρχιτεκτονική του $q_\phi(z_j|r_j)$ είναι παρόμοια με του απλού VAE και παρουσιάζεται στο διάγραμμα 2.14.

$$\begin{aligned}
h_1 &= \texttt{ReLU}(\texttt{FC}_{|\mathbf{r}_j|\to128}(\mathbf{r}_j)) \\
h_2 &= \texttt{ReLU}(\texttt{FC}_{128\to128}(h_1)) \\
\mu &= \texttt{FC}_{128\to128}(h_2) \\
\sigma &= \exp(\tfrac{1}{2}\texttt{FC}_{128\to128}(h_2)) \\
q_\phi(z_j|\mathbf{r}_j) &\sim \mathcal{N}(\mu, \sigma\mathbf{I})
\end{aligned}$$

Σχήμα 2.14: Αρχιτεκτονική του κωδικοποιητή του TVAE

2.6 Maximum Mean Discrepancy

Το Maximum Mean Discrepancy (MMD) είναι ένα στατιστικό τεστ που υπολογίζει την διαφορά μεταξύ των μέσων τιμών των συναρτήσεων δύο δειγμάτων, τα οποία δείγματα ανήκουν στις κατανομές P και Q αντίστοιχα. Αυτή η μέθοδος προτάθηκε από τον Arthur Gretton στην δημοσίευση [13] και έχει σαν σκοπό την εύρεση κατάλληλης συνάρτησης, τέτοια ώστε να ισχύει $MMD = 0$ αν και μόνο αν $P = Q$, αλλά και προϋποθέσεις κάτω από τις οποίες το MMD αποτελεί μετρική.

Το MMD εκμεταλλεύεται το ′κόλπο του πυρήνα′ (kernel trick). Δεδομένης μιας τιμής X, ένα feature map ϕ μεταφέρει το X σε έναν άλλο χώρο $\mathcal{F}$, έτσι ώστε $\phi(X) \in \mathcal{F}$. Θεωρώντας ότι το $\mathcal{F}$ ικανοποιεί τις κατάλληλες προϋποθέσεις (δηλαδή να αποτελεί ένα Reproducing Kernel Hilbert Space ή εν συντομία RKHS όπως περιγράφεται και στο [14]), μπορούμε να χρησιμοποιήσουμε το kernel trick για να υπολογίσουμε το εσωτερικό γινόμενο στον χώρο $\mathcal{F}$:

$$X, Y \text{ τέτοια ώστε } \; k(X, Y) = \langle \phi(X), \phi(Y) \rangle_{\mathcal{F}}$$

Το MMD όπως ειπώθηκε παραπάνω είναι η απόσταση/διαφορά μεταξύ των μέσων τιμών των συναρτήσεων δύο δειγμάτων ή εν συντομία η απόσταση μεταξύ των feature means.

Το εσωτερικό γινόμενο των feature means $X \sim P$ και $Y \sim Q$ γράφεται στην μορφή συνάρτησης πυρήνα ως εξής:

$$\langle \mu_P \ (\phi(X), \mu_Q \ (\phi(Y) \rangle_{\mathcal{F}} = E_{P,Q} \ [\langle \phi(X), \phi(Y) \rangle_{\mathcal{F}}] = E_{P,Q} \ [k(X, Y)]$$

Το MMD εκφράζεται μαθηματικά:

$$MMD^2(P, Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{F}}$$

όπου για λόγους απλότητας παραλείφθηκαν τα $\phi(\cdot)$. Χρησιμοποιώντας την νόρμα $\|x\| = \sqrt{\langle x, x \rangle}$ η προηγούμενη σχέση γίνεται:

$$MMD^2(P, Q) = \langle \mu_P - \mu_Q, \ \mu_P - \mu_Q \rangle = \langle \mu_P, \mu_P \rangle - 2\langle \mu_P, \mu_Q \rangle + \langle \mu_Q, \mu_Q \rangle$$

Χρησιμοποιώντας την αρχική σχέση προκύπτει:

$$MMD^2(P, Q) = E_P \ [k(X, X)] - 2E_{P,Q} \ [k(X, Y)] + E_Q \ [k(Y, Y)]$$

Ως τώρα οι σχέσεις που παρουσιάστηκαν βασίζονται στις κατανομές, όμως σε πραγματικές συνθήκες είναι αρκετά πιθανό να μην είναι γνωστές. Οπότε γίνεται μια εμπειρική εκτίμηση με την παρακάτω σχέση:

$$MMD^2(X, Y) = \frac{1}{m(m-1)} \sum_i \sum_{j \neq i} k(\mathbf{x_i}, \mathbf{x_j}) - 2\frac{1}{m.m} \sum_i \sum_j k(\mathbf{x_i}, \mathbf{y_j}) + \frac{1}{m(m-1)} \sum_i \sum_{j \neq i} k(\mathbf{y_i}, \mathbf{y_j})$$

Μια εξήγηση της σχέσης αυτής είναι ότι τα $\mathbf{x_i}$ αφορούν τα πραγματικά δεδομένα και τα $\mathbf{y_i}$ τα συνθετικά/παραγόμενα δεδομένα, οπότε το MMD αποτελεί μετρική εκτίμησης των μοντέλων της διπλωματικής.

Δύο πολύ συνηθισμένοι πυρήνες (που ικανοποιούν το RKHS) για την χρήση του kernel trick είναι ο πολυωνυμικός $K(x, y) = (x^\mathsf{T} y + c)^d$ και ο γκαουσιανός (γνωστός και ως RBF Kernel) $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, όπου $\gamma = \frac{1}{2\sigma^2}$

2.7 Εργαλεία

2.7.1 nProbe

Για την συλλογή και εξαγωγή των δεδομένων κίνησης του νοσοκομειακού δικτύου χρησιμοποιήθηκε το εργαλείο nProbe, το οποίο παρέχει υποστήριξη για την έκδοση 9 του πρωτοκόλλου NetFlow, η οποία επιλέχθηκε για την παρούσα διπλωματική. Η έκδοση αυτή επιτρέπει την συλλογή αρκετών πληροφοριών για τα δεδομένα ροής δικτύου και με το εργαλείο αυτό δίνει την δυνατότητα για την επιλογή όσων χρειάζονται.

2.7.2 Python

Για την υλοποίηση αυτής της διπλωματικής εργασίας, χρησιμοποιήθηκε εξ ολοκλήρου η προγραμματιστική γλώσσα Python με την βοήθεια αρκετών βιβλιοθηκών της (libraries), εκ των οποίων οι σημαντικότερες παρουσιάζονται παρακάτω.

- **NumPy:** Η NumPy είναι μια χρήσιμη βιβλιοθήκη με εργαλεία για γρήγορους υπολογισμούς μέσω των πολύ αποδοτικών numpy arrays, αλλά και αρκετές υλοποιήσεις μαθηματικών συναρτήσεων.
- **SciPy:** Η SciPy, όπως και η NumPy, προσφέρει πληθώρα εργαλείων και μαθηματικών συναρτήσεων/αλγορίθμων με αποδοτικές υλοποιήσεις.
- **Pandas:** Η Pandas είναι μια βιβλιοθήκη που συμβάλλει στην εύκολη και γρήγορη επεξεργασία και ανάλυση των δεδομένων στην μορφή dataframes.
- **Scikit-Learn:** Μια βιβλιοθήκη που περιέχει εργαλεία για ανάλυση δεδομένων, αλλά και απλούς αλγορίθμους μηχανικής μάθησης (όχι όμως βαθιάς μάθησης).
- **PyTorch:** Βασική βιβλιοθήκη για την διπλωματική αυτή αφού προσφέρει υποστήριξη για μοντέλα βαθιάς μάθησης (όπως τα GANs και οι VAEs).
- **Visualization Packages:** Για την οπτικοποίηση των δεδομένων που προέκυψαν έπειτα από τα στάδια ανάλυσης και παραγωγής, χρησιμοποιήθηκαν τα πακέτα **Matplotlib**, **Seaborn** και **Plotly**, καθένα εκ των οποίων παρείχε τα δικά του πλεονεκτήματα.

Η εκτέλεση των προγραμμάτων για την ανάλυση των δεδομένων και για τα μοντέλα μηχανικής μάθησης, πραγματοποιήθηκε σε περιβάλλοντα όπως το Jupyter Notebook, το Google Colaboratory και το Kaggle, εκ των οποίων τα δύο τελευταία παρέχουν και υποστήριξη με GPU, που ήταν αρκετά χρήσιμη, ιδιαίτερα για τα μοντέλα μηχανικής μάθησης (λόγω του μεγάλου όγκου δεδομένων κατά την εκπαίδευση).

Κεφάλαιο 3

Διερευνητική Ανάλυση Δεδομένων

3.1 Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπόνηση της παρούσας διπλωματικής, προέρχεται από νοσοκομειακό περιβάλλον (που για λόγους ασφάλειας προσωπικών δεδομένων δεν αναφέρεται στην διπλωματική) και συλλέχθηκε από υπεύθυνους του **Τομέα Ηλεκτρονικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων** σε συνεργασία με υπεύθυνους του νοσοκομείου, με χρήση του εργαλείου **nProbe**.

Τα δεδομένα αποθηκεύτηκαν σε μορφή **NetFlow** κίνησης (ένα αρχείο ανά λεπτό) για την χρονική περίοδο ***2021/04/07-2021/05/09***. Τα πεδία που καταγράφηκαν φαίνονται αναλυτικά στον πίνακα 3.1 με επεξήγηση για τα περιεχόμενα κάθε πεδίου.

3.1.1 Επεξεργασία/Τροποποίηση Δεδομένων

Πριν παραδοθούν αυτά τα δεδομένα στους υπεύθυνους του Τομέα, ήταν επιτακτική η ανάγκη για την απόκρυψη ευαίσθητων προσωπικών δεδομένων του νοσοκομειακού προσωπικού. Οπότε για τον λόγο αυτό, πραγματοποιήθηκαν σημαντικές αλλαγές στα δεδομένα, προκειμένου να διατηρηθεί όσο το δυνατόν περισσότερη πληροφορία πριν αφαιρεθούν τα επίμαχα πεδία (όπως **IPv4 Source Address** και **HTTP URL**).

Οι ενέργειες που πραγματοποιήθηκαν για τον μετασχηματισμό των δεδομένων, περιγράφονται παρακάτω:

- Οι Source IP's αντικαταστάθηκαν με μια στήλη από unique ID's, για να διατηρηθεί η μοναδικότητα της κάθε διεύθυνσης, και με μια προσωρινή στήλη που περιέχει γενικές πληροφορίες σχετικά με τον τομέα στον οποίο ανήκουν.

- Για να αποφευχθεί το πρόβλημα των πολλαπλών IP ανά συσκευή εξαιτίας των lease renewals (που αναφέρθηκαν στο 2 στην ενότητα του DHCP), χρησιμοποιήθηκαν τα DHCP Logs που είχαν καταγραφεί από τους υπευθύνους του νοσοκομείου, ώστε κάθε IP να αντιστοιχιστεί μόνο με μια συσκευή σε κάθε χρονική στιγμή.

- Η στήλη HTTP URL χρησιμοποιήθηκε για να εξαχθούν οι ιστοσελίδες που επισκέφθηκε ο χρήστης (π.χ. Google, Youtube, κ.α.), χωρίς επιπλέον πληροφορίες που πιθανώς να περιείχαν ευαίσθητα δεδομένα.

NetFlow v9 Field	Description
IPV4_SRC_ADDR	IPv4 source address
IPV4_DST_ADDR	IPv4 destination address
IN_SRC_MAC	Source MAC Address
OUT_DST_MAC	Post Destination MAC Address
IN_DST_MAC	Destination MAC Address
OUT_SRC_MAC	Post Source MAC Address
L4_SRC_PORT	IPv4 source port
L4_DST_PORT	IPv4 destination port
PROTOCOL	IP protocol byte
L7_PROTO	Layer 7 protocol (numeric)
L7_PROTO_NAME	Layer 7 protocol name
L7_PROTO_CATEGORY	Layer 7 protocol category
IN_BYTES	Incoming flow bytes ($src \rightarrow dst$)
IN_PKTS	Incoming flow packets ($src \rightarrow dst$)
OUT_BYTES	Outgoing flow bytes ($dst \rightarrow src$)
OUT_PKTS	Outgoing flow packets ($dst \rightarrow src$)
FLOW_END_REASON	The reason for flow termination
TCP_FLAGS	Cumulative of all flow TCP flags
SERVER_TCP_FLAGS	Cumulative of all server TCP flags
DIRECTION	Flow direction [0=Receive, 1=Transmit]
FLOW_START_MILLISECONDS	Msec (epoch) of the first flow packet
FLOW_END_MILLISECONDS	Msec (epoch) of the last flow packet
ICMP_TYPE	ICMP Type * 256 + ICMP code
HTTP_URL	HTTP URL (IXIA URI)
HTTP_METHOD	HTTP method
HTTP_RET_CODE	HTTP return code (e.g. 200, 304...)
DNS_QUERY	DNS query
DNS_QUERY_TYPE	DNS query type (e.g. 1=A, 2=NS..)
DNS_RET_CODE	DNS return code (e.g. 0=no error)
DNS_RESPONSE	DNS response(s)

Πίνακας 3.1: Πεδία που καταγράφηκαν αρχικά με το εργαλείο nProbe

- Επιπλέον πληροφορίες από τους υπευθύνους, όπως ο τομέας στον οποίο ανήκει η συσκευή προέλευσης ή προορισμού της κίνησης, αποθηκεύτηκαν σε δύο νέα πεδία με όνοματα SRC_MACHINE, DST_MACHINE. Παραδείγματα τομέων είναι τα επείγοντα, το καρδιολογικό, η μισθοδοσία προσωπικού κ.α.

- Δημιουργήθηκε νέο πεδίο με όνομα 'Traffic' για ειδική κίνηση (π.χ. DICOM, BMS, LIS).

- Η κατηγορία 'Server' στα πεδία SRC_MACHINE, DST_MACHINE αναφέρεται στην κίνηση συσκευών τύπου DNS, Secondary DNS, DHCP, Active Directory, NAS για λόγους απλότητας μετά από υπόδειξη των υπευθύνων του νοσοκομείου.

- Οι συσκευές του δικτύου που δεν συσχετίστηκαν με κάποιο τομέα του νοσοκομείου (όπως περιγράφηκε παραπάνω), ονοματίστηκαν 'User Desktop' ή 'Mobile' ή 'Network Device', σύμφωνα με επιπλέον πληροφορίες που δόθηκαν από τους υπεύθυνους του νοσοκομείου. Ουσιαστικά είναι κάτι σαν μια γενική κλάση, αυτή που περιέχει τις συσκευές που δεν ανήκουν σε κάποιον τομέα (π.χ. είναι προσωπικές συσκευές όπως τα κινητά) ή δεν υπάρχει γνώση της προέλευσης τους.

3.2 Ανάλυση Δεδομένων

Αφότου παραλήφθηκαν τα τροποποιημένα δεδομένα, πραγματοποιήθηκε Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis ή EDA εν συντομία) σε αυτά, αρχικά για την εκτίμηση των σημαντικών πεδίων του συνόλου δεδομένων και εν συνεχεία για την μελέτη πιο σύνθετων σχέσεων μεταξύ των δεδομένων.

3.2.1 Μελέτη όγκου ροής δεδομένων με χρονοσειρές

3.2.1.1 Ανά ημέρα

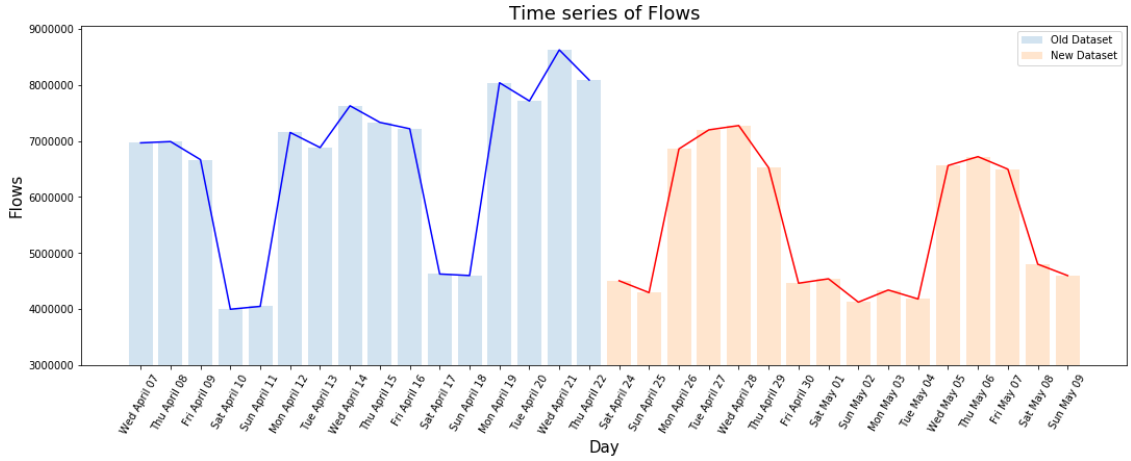
Αρχικά για να εκτιμηθεί η κίνηση των νοσοκομειακών χρηστών, μελετήθηκαν οι χρονοσειρές ροής δικτύου, bytes και πακέτων ανά ημέρα, οι οποίες παρουσιάζονται στα διαγράμματα 3.1, 3.2, 3.3.

Από αυτά τα διαγράμματα χρονοσειρών μπορούν να εξαχθούν τα εξής συμπεράσματα:

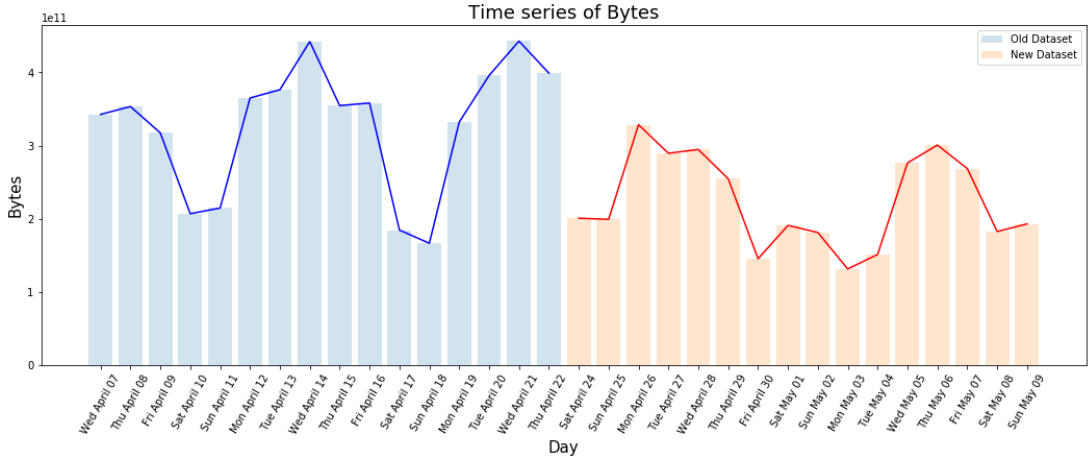
- Τα διαγράμματα χρονοσειρών bytes και πακέτων είναι πανομοιότυπα και αυτό είναι λογικό αφού συνηθίζεται να έχουν αναλογική σχέση.

- Τα Σαββατοκύριακα, η κίνηση των χρηστών είναι εμφανώς μειωμένη, κάτι που παρατηρείται και στα τρία διαγράμματα και είναι αναμενόμενο, διότι οι περισσότεροι νοσοκομειακοί και ιδιαίτερα το διοικητικό προσωπικό, δεν εργάζονται αυτές τις μέρες (με εξαίρεση τομείς όπως τα επείγοντα)

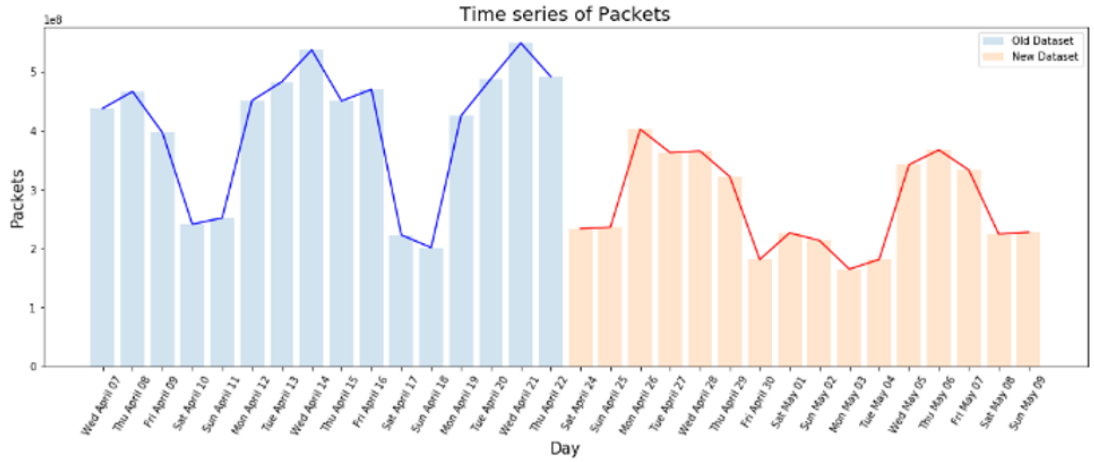
- Αντίστοιχη συμπεριφορά με τα Σαββατοκύριακα παρατηρείται επιπλέον και στις ημερομηνίες 26/04/2021 - 02/05/2021, 03/05/2021 και 04/05/2021, όπου είναι επίσημες εθνικές αργίες (εβδομάδα του Πάσχα, Καθαρά Δευτέρα και Εργατική Πρωτομαγιά αντίστοιχα). Η κίνηση σε αυτές τις μέρες είναι σημαντικά μειωμένη συγκριτικά με τις αντίστοιχες ημέρες των άλλων εβδομάδων. Η επιπρόσθετη αυτή μείωση πιθανώς οφείλεται στις άδειες που δίνονται στο προσωπικό για αυτές τις μέρες, σε συνδυασμό με τον μειωμένο αριθμό ιατρικών ραντεβού (πάλι λόγω των ημερών).



Σχήμα 3.1: Συσσωρευτικός αριθμός ροής δικτύου ανά ημέρα



Σχήμα 3.2: Συσσωρευτικός αριθμός bytes ανά ημέρα

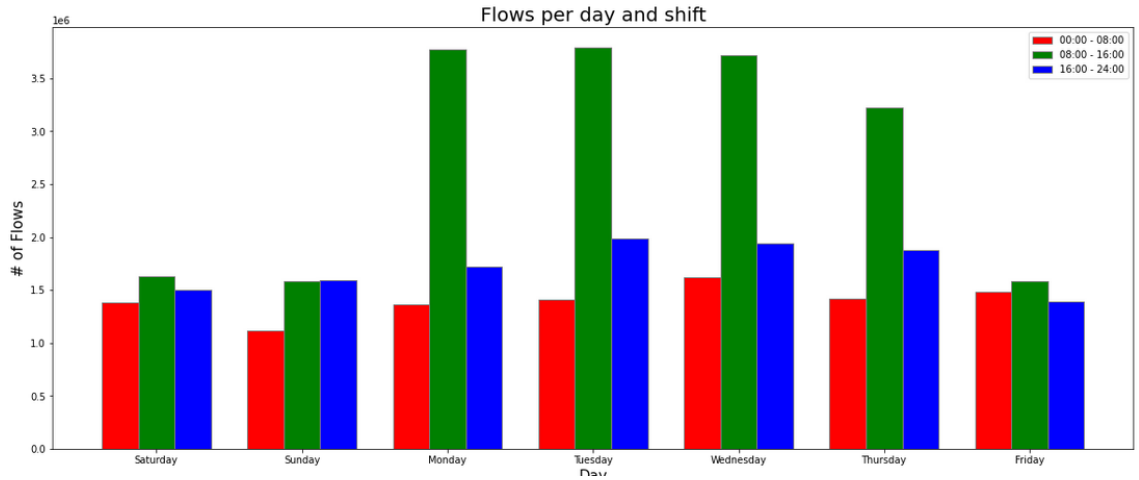


Σχήμα 3.3: Συσσωρευτικός αριθμός πακέτων ανά ημέρα

Να σημειωθεί ότι η 23η Απριλίου λείπει από τις χρονοσειρές διότι η συλλογή των δεδομένων είχε σταματήσει για σύντομο χρονικό διάστημα εκείνη την ημέρα, γεγονός που θα καθιστούσε την συμπερίληψη της λανθασμένη. Για αυτό τον λόγο και στα παραπάνω διαγράμματα υπάρχει η επισημείωση Old Dataset και New Dataset, που επισημαίνει ποια δεδομένα προήλθαν πριν την διακοπή και ποια μετά.

3.2.1.2 Ανά βάρδια

Για την περαιτέρω εκτίμηση της κίνησης, αποφασίστηκε η μελέτη όχι μόνο ανά ημέρα, αλλά και ανά βάρδια, για την εβδομάδα 24/04/2021 - 30/04/2021. Οι 3 βάρδιες που μελετήθηκαν είναι οι 00:00-08:00, 08:00-16:00 και 16:00-24:00 και το διάγραμμα ροής δικτύου φαίνεται παρακάτω 3.4.



Σχήμα 3.4: Συσσωρευτικός αριθμός ροής δικτύου ανά βάρδια για την εβδομάδα 24/04/2021-30/04/2021

Από το αυτό το διάγραμμα οι νέες παρατηρήσεις είναι:

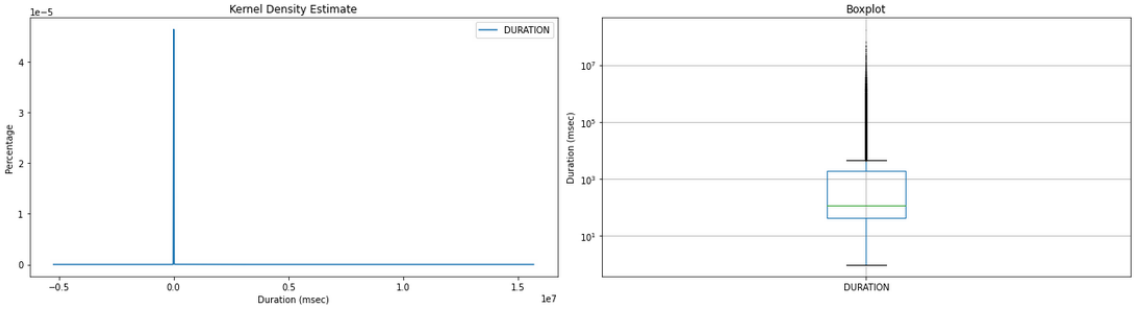
- Η βάρδια 08:00-16:00 έχει την περισσότερη κίνηση από τις τρεις και η βάρδια 16:00-00:00 την αμέσως επόμενη. Αυτό οφείλεται στο γεγονός ότι η πλειονότητα των διοικητικών υπαλλήλων του νοσοκομείου εργάζεται τις πρωινές/μεσημεριανές ώρες, σε αντίθεση με το ιατρικό προσωπικό που έχει σταθερή παρουσία σε όλες τις βάρδιες κάτι που μας οδηγεί στα επόμενα συμπεράσματα.

- Τις καθημερινές η κίνηση είναι περισσότερη από το Σαββατοκύριακο και επίσης η κίνηση το Σαββατοκύριακο είναι σχεδόν ισομοιρασμένη μεταξύ των τριων βαρδιών. Αυτό αποδεικνύει και τον παραπάνω ισχυρισμό, για το ιατρικό προσωπικό, αφού οι περισσότεροι διοικητικοί υπάλληλοι δεν εργάζονται τα Σαββατοκύριακα.

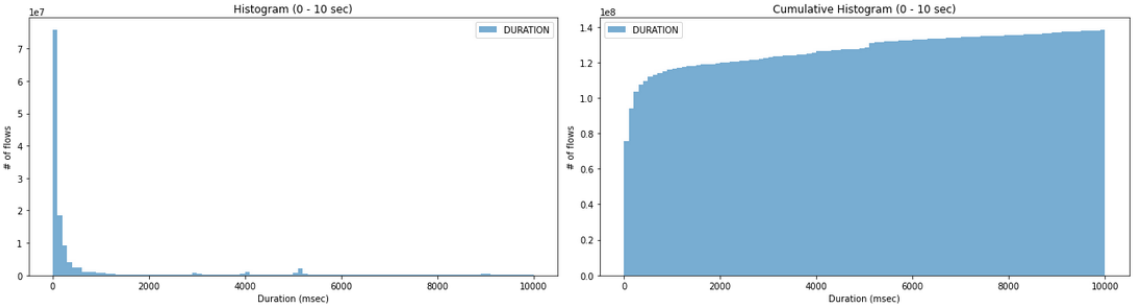
- Η Παρασκευή φαίνεται σαν εξαίρεση στον κανόνα (όσον αφορά τις καθημερινές), αλλά δεν είναι μια τυχαία μέρα. Πρόκειται για την Μεγάλη Παρασκευή και όπως είναι φυσιολογικό, η κίνηση είναι σημαντικά μειωμένη.

3.2.2 Μελέτη διάρκειας ροής

Στο διάγραμμα 3.5α′ φαίνεται η κατανομή της διάρκειας ροής όλων των ροών της βάσης δεδομένων μέσω ενός διαγράμματος πυκνότητας πυρήνα και ενός boxplot.



(α′) Διάγραμμα εκτίμησης πυκνότητας πυρήνα και boxplot διάρκειας ροής



(β′) Ιστόγραμμα και αθροιστικό ιστόγραμμα για διάρκεια ροής έως 10 δευτερόλεπτα

Σχήμα 3.5: Διαγράμματα διάρκειας ροής

Από την εκτίμηση πυκνότητας πυρήνα δεν μπορούν να εξαχθούν πολλά συμπεράσματα γιατί έχει μια ραγδαία αύξηση κοντά στην τιμή 0. Από το boxplot όμως φαίνεται πιο καθαρά ότι η πλειονότητα των διαρκειών ροής βρίσκονται ανάμεσα στα 0.1 με 10 δευτερόλεπτα.

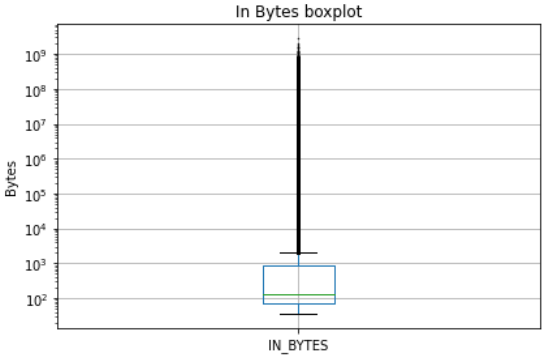
Να σημειωθεί ότι το boxplot είναι ένα διάγραμμα του οποίου το κουτί (box) συμβολίζει τις τιμές μεταξύ 25ου και 75ου εκατοστημορίου (οι οποίες είναι και οι πιο συνήθεις) και το διάστημα αυτό ονομάζεται διατεταρτημοριακό εύρος (interquantile range, IQR). Οι τιμές μεταξύ των γραμμών είναι το εύρος $(Q1 - 1.5 * IQR, Q3 + 1.5 * IQR)$ και οι τιμές εκτός αυτού θεωρούνται ακραίες τιμές.

Το γεγονός ότι οι περισσότερες διάρκειες ροής είναι τόσο μικρές/σύντομες οφείλεται στο ότι πολλά δεδομένα ροής δικτύου δημιουργούνται από μόνα τους και όχι από επιλογή του χρήστη (για παράδειγμα αν ένας χρήστης ανοίξει μια ιστοσελίδα, εκτός από την κύρια ροή, δημιουργούνται και αρκετές μικρότερες, όπως αιτήματα DNS, αναδυόμενα παράθυρα κ.α. Το ίδιο συμβαίνει και σε βίντεο στο Youtube και άλλες εφαρμογές).

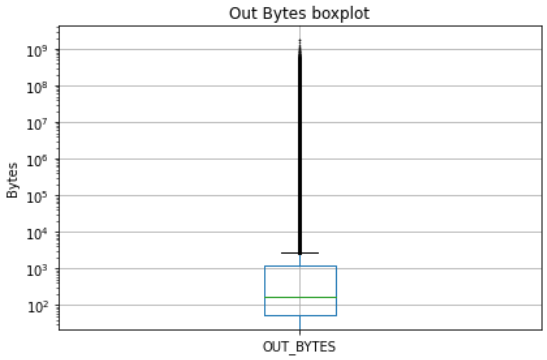
Αυτό φαίνεται και στα ιστογράμματα του σχήματος 3.5β′, τα οποία μελετάνε τις διάρκειες ροής από 0 έως 10 δευτερόλεπτα (που αποτελούν το 87% του συνόλου των δεδομένων).

3.2.3 Μελέτη bytes και πακέτων ροής

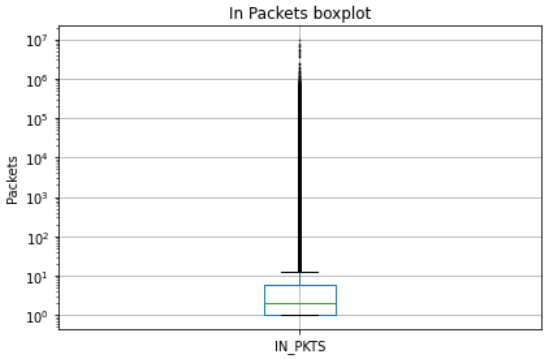
Στο διάγραμμα 3.6 φαίνονται σε ποια εύρη κατανέμονται οι τιμές των εισερχόμενων και εξερχόμενων bytes και πακέτων ροής μέσω των τεσσάρων boxplots.



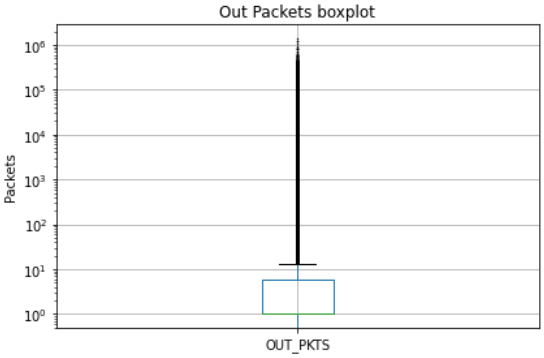
(α′) Boxplot εισερχόμενων bytes



(β′) Boxplot εξερχόμενων bytes



(γ′) Boxplot εισερχόμενων πακέτων



(δ′) Boxplot εξερχόμενων πακέτων

Σχήμα 3.6: Boxplots bytes και πακέτων

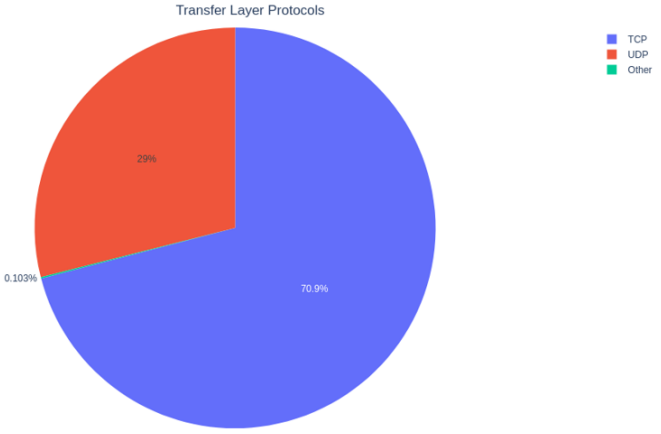
Σαν απόρροια της διάρκειας ροής, είναι λογικό που και τα bytes και τα πακέτα ακολουθούν αντίστοιχη συμπεριφορά, δηλαδή τα περισσότερα συγκεντρώνονται σε χαμηλές τιμές και οι υψηλές (που αντιστοιχούν σε λίγα δεδομένα ροής) θεωρούνται ακραίες.

Στα παραπάνω διαγράμματα παρατηρείται ότι οι συνηθισμένες τιμές των bytes είναι λίγο λιγότερο από 100 έως 1000, ενώ όσον αφορά τα πακέτα οι τιμές αυτές είναι 1 έως 10.

Παρ΄ όλα αυτά σε αντίθεση με το boxplot της διάρκειας ροής, τα διαγράμματα αυτά δείχνουν ότι υπάρχουν αρκετά outliers (από το πόσο πυκνά φαίνονται τα bullets) και όχι μόνο αυτό, αλλά φτάνουν και σε πολύ μεγάλες τιμές. Πιο συγκεκριμένα, τα εισερχόμενα και εξερχόμενα bytes φτάνουν έως και 10^9, τα εισερχόμενα πακέτα έως και 10^7, ενώ τα εξερχόμενα πακέτα έως 10^6.

3.2.4 Μελέτη κατανομής του πρωτοκόλλου επιπέδου μεταφοράς

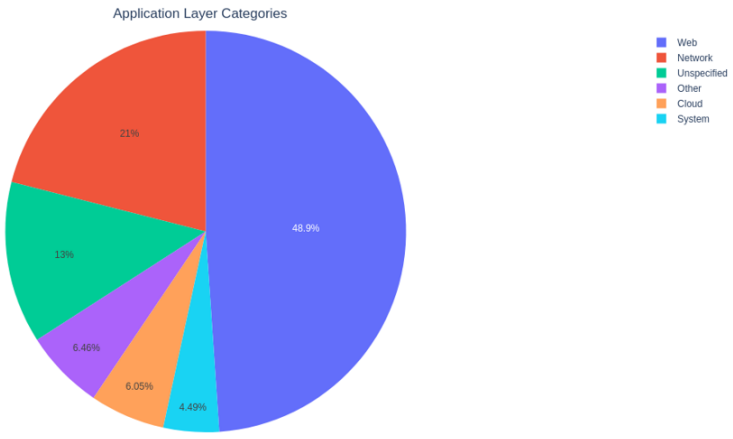
Στο διάγραμμα 3.7 απεικονίζεται η κατανομή των διαφόρων πρωτοκόλλων επιπέδου μεταφοράς (transfer layer protocols). Όπως ήταν αναμενόμενο το βασικό πρωτόκολλο είναι το TCP (Transmission Control Protocol) που είναι ευρέως χρησιμοποιούμενο για την ανταλλαγή δεδομένων μεταξύ επικοινωνιών. Βέβαια μεγάλο ποσοστό (αλλά σαφώς μικρότερο του TCP) καταλαμβάνει και το πρωτόκολλο UDP (User Datagram Protocol) αφού χρησιμοποιείται ακόμη (παρά τα μειονεκτήματα του όσον αφορά την αξιοπιστία) σε εφαρμογές προβολής βίντεο/μουσικής ή σε DNS queries.



Σχήμα 3.7: Διάγραμμα πίτας για τα πρωτόκολλα επιπέδου μεταφοράς

3.2.5 Μελέτη κατανομής των κατηγοριών επιπέδου εφαρμογής

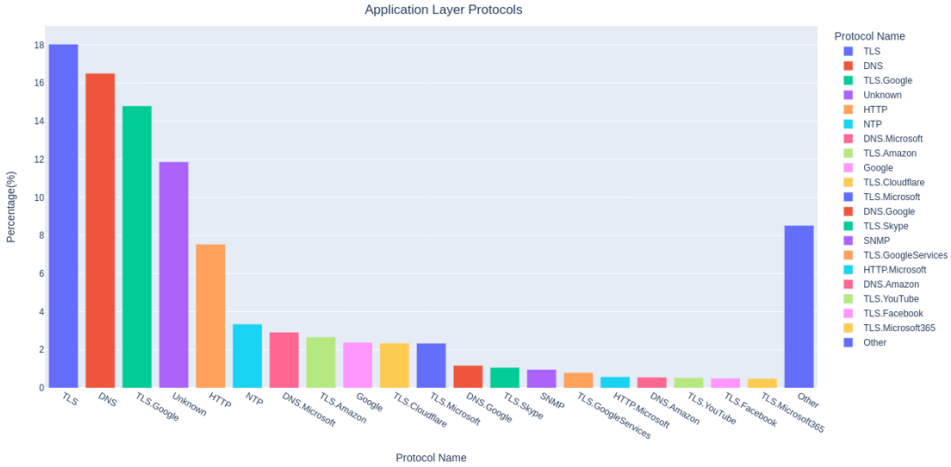
Στο διάγραμμα 3.8 παρουσιάζονται οι κατηγορίες του L7 πρωτοκόλλου. Τα web based πρωτόκολλα απαρτίζουν σχεδόν το 50% της κίνησης ενώ ακολουθούν κατηγορίες όπως network και cloud. Υπάρχει και ένα 13% απροσδιόριστο, που πιθανώς αδυνατεί να καταγράψει το εργαλείο nProbe.



Σχήμα 3.8: Διάγραμμα πίτας για τις κατηγορίες επιπέδου εφαρμογής

3.2.6 Μελέτη κατανομής του πρωτοκόλλου επιπέδου εφαρμογής

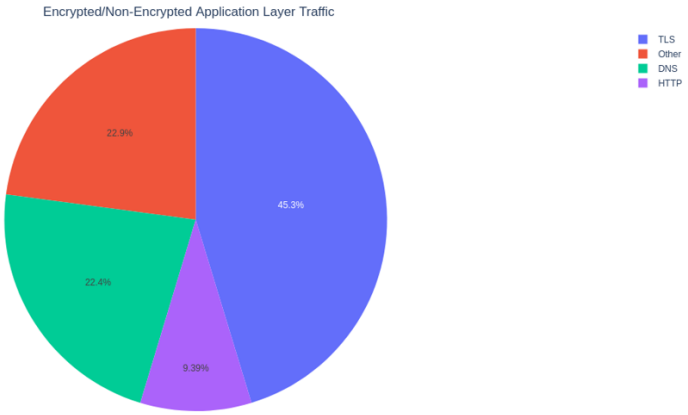
Εξετάζοντας τα πρωτόκολλα που συνέστησαν τις παραπάνω κατηγορίες, προέκυψε το διάγραμμα 3.9, στο οποίο φαίνεται ξεκάθαρα ότι η πλειονότητα της ροής δικτύου χρησιμοποιεί το πρωτόκολλο κρυπτογράφησης TLS ακολουθούμενο από τα πρωτόκολλα DNS και HTTP. Επίσης εμφανίζονται και κάποια δευτερεύοντα ονόματα που υποδηλώνουν την υπηρεσία που χρησιμοποιήθηκε για να παραχθεί αυτή η ροή, με σημαντικότερες τις Google, Amazon και Microsoft.



Σχήμα 3.9: Διάγραμμα μπαρών για τα πρωτόκολλα επιπέδου εφαρμογής (μαζί με δευτερεύουσα πληροφορία)

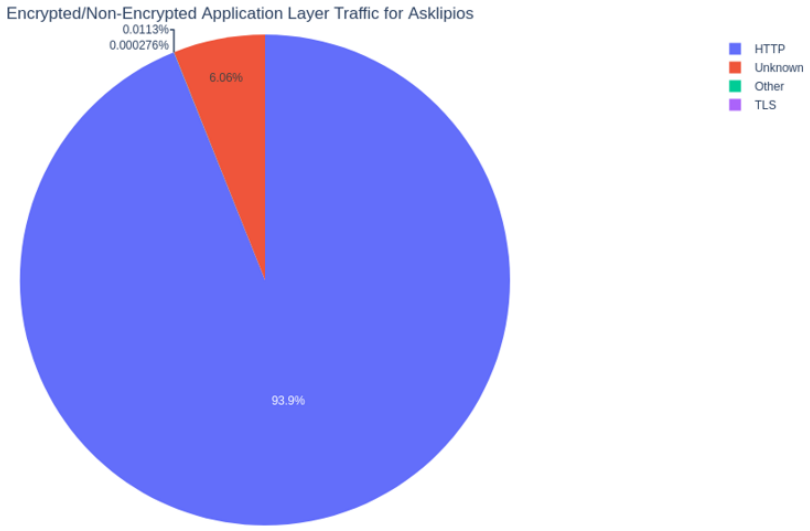
3.2.7 Ανάλυση κρυπτογράφησης της ροής του HIS

Για να αναλυθούν τα πρωτόκολλα κρυπτογράφησης που διέπουν την κίνηση του HIS (Hospital Information Systems), αρχικά ενοποιήθηκαν τα πρωτόκολλα του παραπάνω διαγράμματος, ώστε να αφαιρεθεί η πληροφορία της υπηρεσίας. Το διάγραμμα 3.10 που προέκυψε δείχνει ότι σε όλη την κίνηση η αναλογία κρυπτογράφησης-μη κρυπτογράφησης είναι σχεδόν 5 προς 1 (45% έναντι 9%).



Σχήμα 3.10: Διάγραμμα πίτας για τα πρωτόκολλα επιπέδου εφαρμογής

Στην βασικότερη όμως ιατρική υπηρεσία, το HIS (που στο διάγραμμα 3.11 εμφανίζεται και σαν Ασκληπιός), φαίνεται ξεκάθαρα ότι δεν υπάρχει σχεδόν καθόλου κρυπτογραφημένη κίνηση, με σχεδόν το 95% της κίνησης να χρησιμοποιεί πρωτόκολλο HTTP (δηλαδή χωρίς κρυπτογράφηση) και το εναπομείναν ποσοστό να είναι άγνωστο τι πρωτόκολλο χρησιμοποιεί. Αυτό επιβεβαιώνει τις ανησυχίες που περιγράφηκαν στην εισαγωγή αυτής της διπλωματικής σχετικά με την κυβερνοασφάλεια και δημιουργεί επιτακτική ανάγκη εκτός από την παρακολούθηση της ροής και την αναβάθμιση του πρωτοκόλλου ασφαλείας. Να σημειωθεί επίσης ότι η HTTP-κίνηση του HIS αποτελεί το 30% της συνολικής HTTP-κίνησης (4.76 εκατομμύρια flows στα 15.73 εκατομμύρια).



Σχήμα 3.11: Διάγραμμα πίτας για τα πρωτόκολλα επιπέδου εφαρμογής του HIS

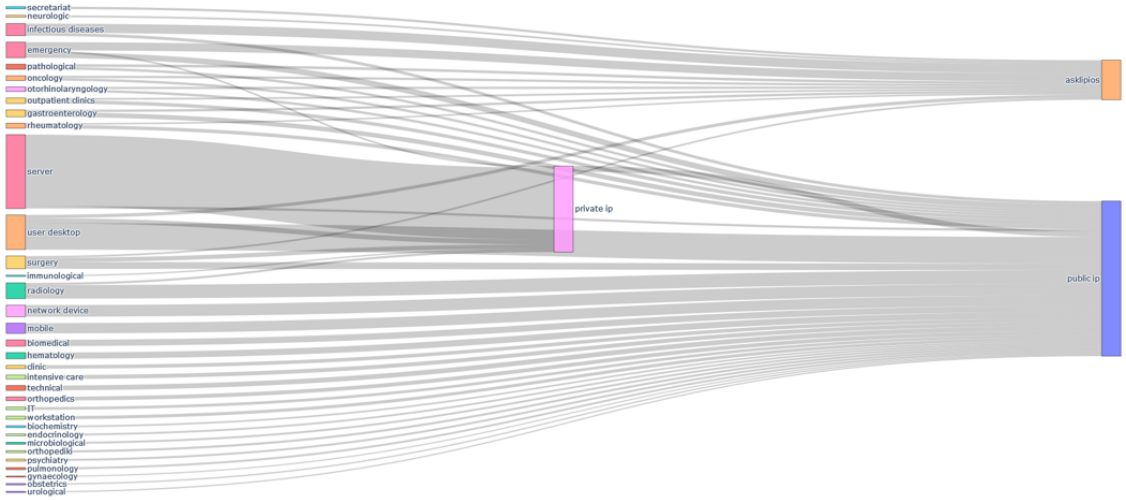
3.2.8 Παρακολούθηση συμπεριφοράς χρηστών/συσκευών διαφόρων τομέων

Σε αυτή την ενότητα μελετώνται τα πεδία SRC_MACHINE και DST_MACHINE. Για να γίνει καλύτερη οπτικοποίηση κάποιων διαγραμμάτων αποφασίστηκε η δημιουργία κάποιων γενικών κατηγοριών για τα πεδία αυτά. Οι κατηγορίες είναι:

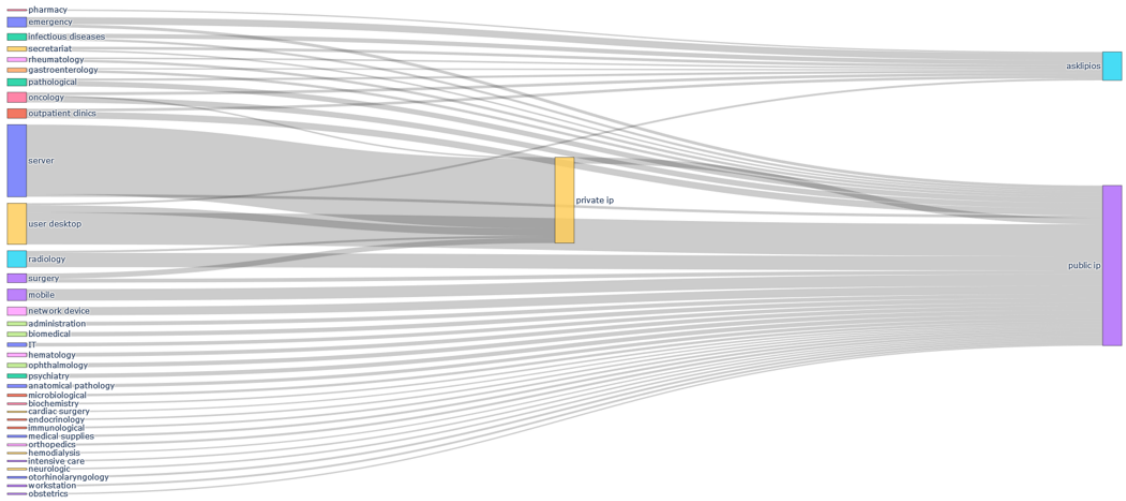
- **Public IP:** Εμπεριέχει κάθε διεύθυνση IP που είναι ορατή στο διαδίκτυο όπως τα Google, Youtube, Facebook κ.α. Εξαίρεση αποτελεί η IP του promitheus.gov η οποία αν και δημόσια, επιλέχθηκε να εξεταστεί σαν ξεχωριστή κατηγορία.
- **Private IP:** Εμπεριέχει κάθε διεύθυνση IP που χρησιμοποιείται σε εσωτερικά δίκτυα (όπως του νοσοκομείου). Εξαίρεση εδώ αποτελεί η IP του Ασκληπιού (που αναφέρεται ως HIS στην παρούσα διπλωματική) που ενώ είναι ιδιωτική, διαχωρίστηκε για να εξεταστεί ξεχωριστά.
- **Server:** Σε αυτή την κατηγορία υπάγονται διάφοροι servers όπως DHCP, DNS, NAS και Active Directory του νοσοκομείου.

3.2.8.1 Ποιοτικά χαρακτηριστικά

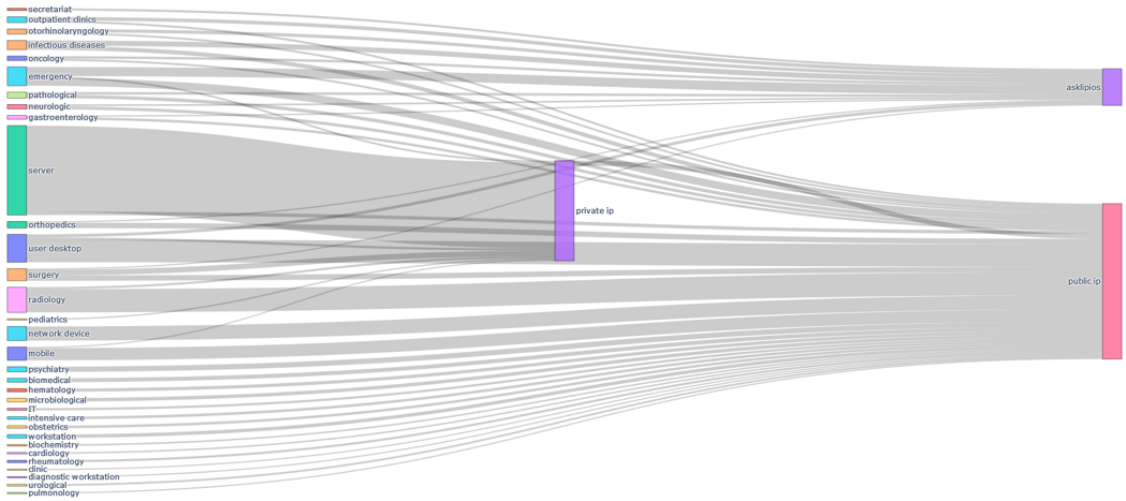
Για την οπτικοποίηση των σχέσεων μεταξύ των πεδίων SRC_MACHINE, DST_MACHINE χρησιμοποιήθηκαν διαγράμματα Sankey, που αποτελούν κατηγορία των διαγραμμάτων ροής και στα οποία το πλάτος των βελών είναι ανάλογο του ποσοστού ροής. Επειδή όμως όλα τα ζεύγη SRC_MACHINE, DST_MACHINE ήταν αδύνατον να οπτικοποιηθούν (ήταν περισσότερα από 10000, με την μειονότητα αυτών να έχουν την πλειονότητα των εμφανίσεων, που επιβεβαιώνει και την θεμελιώδη αρχή του Pareto [15]), επιλέχθηκαν τα 50 με τις περισσότερες εμφανίσεις. Οι μέρες στις οποίες πραγματοποιήθηκε η ανάλυση είναι η Τετάρτη 24 Απριλίου, το Σάββατο 28 Απριλίου και η Τρίτη 4 Μαΐου. Ο λόγος για την επιλογή των τριών αυτών ημερών είναι ότι καθεμία αποτελεί διαφορετική κατηγορία ημέρας, δηλαδή καθημερινή, Σαββατοκύριακο και εθνική αργία (όπως είχε προκύψει και στην πρώτη ενότητα αυτού του κεφαλαίου). Τα αποτελέσματα παρουσιάζονται στα διαγράμματα 3.12, 3.13 και 3.14.



Σχήμα 3.12: Διάγραμμα Sankey των δυάδων SRC_MACHINE, DST_MACHINE το Σάββατο 24 Απριλίου



Σχήμα 3.13: Διάγραμμα Sankey των δυάδων SRC_MACHINE, DST_MACHINE την Τετάρτη 28 Απριλίου



Σχήμα 3.14: Διάγραμμα Sankey των δυάδων SRC_MACHINE, DST_MACHINE την Τρίτη 4 Μαΐου (Εθνική Αργία)

Πριν παρουσιαστούν τα συμπεράσματα που απορρέουν από τα παραπάνω διάγραμματα, να σημειωθεί ότι η παρουσία του 'Server' σαν SRC_MACHINE είναι περίεργη και κατά πάσα πιθανότητα λανθασμένη, αφού συνήθως οι συσκευές του δικτύου είναι αυτές που ξεκινούν την σύνδεση με έναν server και όχι το αντίστροφο, όπως προκύπτει από τα διαγράμματα. Αυτό το σφάλμα λογικά οφείλεται στην αδυναμία του εργαλείου nProbe να καταγράψει σωστά αυτές τις συνδέσεις και γι' αυτό δεν θα εξεταστεί περαιτέρω η συγκεκριμένη κατηγορία στις εναπομείνασες υποενότητες.

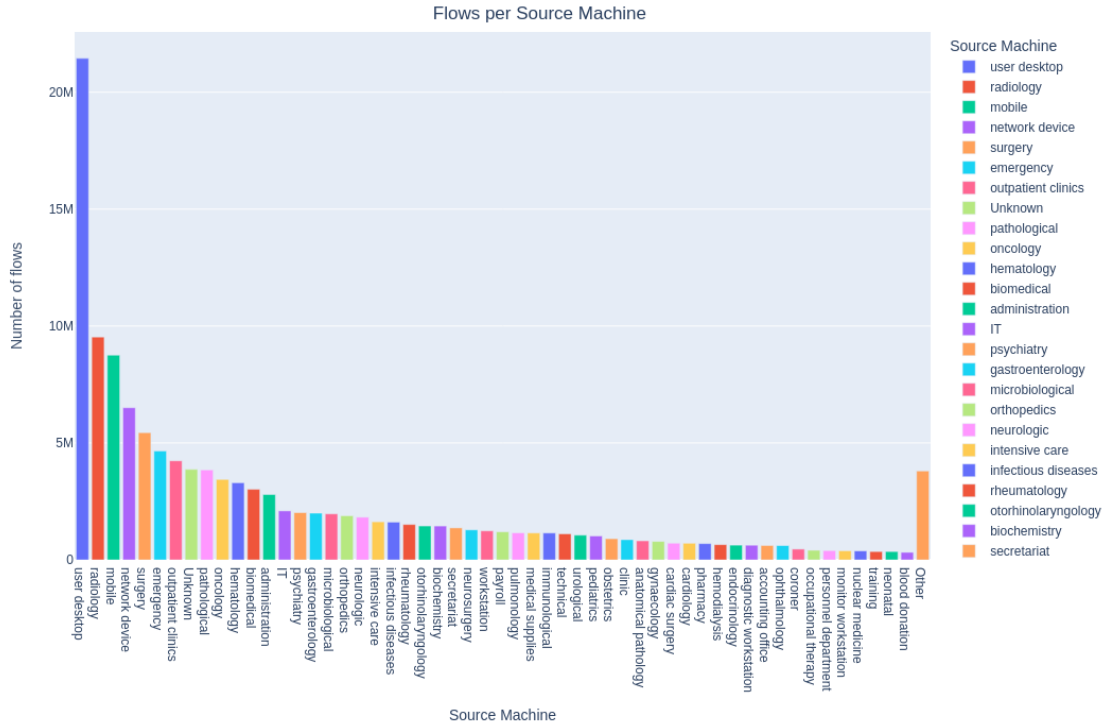
Τα συμπεράσματα μπορούν να συνοψιστούν ως εξής:

- Η συνολική κίνηση το Σάββατο και την ημέρα εθνικής αργία είναι χαμηλότερη απ' ότι της Τετάρτης που είναι καθημερινή, αλλά η κίνηση από ορισμένους ιατρικούς τομείς είναι σχεδόν ίδια, ίσως και περισσότερη (όπως για παράδειγμα τα επείγοντα και το ακτινολογικό προς την υπηρεσία του HIS). Η μείωση κίνησης οφείλεται κυρίως στην μείωση προς τις δημόσιες IP's που γίνεται από προσωπικούς υπολογιστές και κινητά (τα οποία μπορούν να ανήκουν τόσο σε ιατρικό όσο και διοικητικό προσωπικό).

- Είναι επίσης άξιο να αναφερθεί ότι συγκεκριμένες δυάδες εμφανίζονται και στις τρεις μέρες και όλες αφορούν κίνηση από ιατρικό τομέα (επείγοντα, λοιμωδών, ογκολογικό, εξωτερικά ιατρεία κ.α.) προς την υπηρεσία του HIS και άλλες δημόσιες IP's. Αυτό ενισχύει περαιτέρω τα προηγούμενα συμπεράσματα αυτού του κεφαλαίου σχετικά με την εμφάνιση των διοικητικών κυρίως στην βάρδια 08:00-16:00 που φαίνεται και στην μελέτη της κίνησης που πραγματοποιήθηκε ανά βάρδια.

3.2.8.2 Ποσοτικά χαρακτηριστικά

Στο διάγραμμα 3.15 εμφανίζεται ο αριθμός των δεδομένων ροής ανά SRC_MACHINE. Συνολικά υπάρχουν 91 ξεχωριστές κατηγορικές τιμές σε αυτό το πεδίο και κρατήθηκαν οι πρώτες 55 για λόγους οπτικοποίησης. Επίσης η τιμή 'Server' δεν υπολογίστηκε στα αποτελέσματα για τους λόγους που προαναφέρθηκαν, αλλά αξίζει να αναφερθεί ότι είχε υπερδιπλάσια κίνηση από την αμέσως επόμενη τιμή που είναι το 'User Desktop'. Οι υπόλοιπες 36 τιμές ομαδοποιήθηκαν και αποτελούν λιγότερο από το 4% της συνολικής κίνησης.



Σχήμα 3.15: Αριθμός δεδομένων ροής ανά SRC_MACHINE

Συμπεραίνοντας, με εξαίρεση τις τιμές 'User Desktop', 'Mobile' και 'Network Device' που δεν υποδηλώνουν κάποιον τομέα του νοσοκομείου, οι κυριότερες τιμές που εμφανίζονται αφορούν ιατρικούς τομείς και λιγότερο διοικητικούς, κάτι που είναι λογικό αν αναλογιστεί κανείς ότι οι ιατρικοί τομείς είναι επανδρωμένοι όλο το 24ωρο, ενώ οι διοικητικοί συνήθως για ένα 8ωρο. Σημαντικοί τέτοιο τομείς είναι το ακτινολογικό (radiology), το χειρουργικό (surgery), τα επείγοντα (emergency) και τα εξωτερικά ιατρεία (outpatient clinics).

3.3 Κατηγοριοποίηση Χρηστών

Για την επίτευξη του σταδίου της παραγωγής δεδομένων, χρειάστηκε η περαιτέρω μοντελοποίηση των δεδομένων με την δημιουργία ενός νέου πεδίου που θα περιέχει τις κλάσεις συγκεκριμένων χρηστών. Αυτοί οι χρήστες (που αριθμούν 707) επιλέχθηκαν λόγω της κίνησης τους προς συγκεκριμένες υπηρεσίες (ιατρικού ή διοικητικού σκοπού), οι οποίες φαίνονται στον πίνακα 3.2.

Με την προσθήκη αυτού του πεδίου δίνεται η δυνατότητα για την ομαδοποίηση χρηστών με παρεμφερή συμπεριφορά, ώστε να παραχθούν συμπαγή προφίλ που θα βοηθήσουν στην παραγωγή δεδομένων (π.χ. αν αποφασιστεί η παραγωγή δεδομένων ροής δικτύου από γιατρό

που μπαίνει σε κάποια επιλεγμένη υπηρεσία, να μπορεί να προσομοιωθεί με τα υπάρχοντα δεδομένα γιατρών προς αυτή την υπηρεσία). Δεν είναι δηλαδή σκοπός η παραγωγή δεδομένων με βάση τα δεδομένα ενός μόνο χρήστη, αλλά με βάση όλους τους χρήστες που ανήκουν στην ίδια ομάδα (που θα εξηγηθεί επαρκώς στο κεφάλαιο 5 πως προκύπτουν οι ομάδες και από ποιους παράγοντες επηρεάζονται).

User Categories / Target Services	Doctor	Admin	Nurse	Pharmacist	Secretariat
HIS (Port 51001)	✓		✓		
HIS (Port 7778)		✓			
Promitheus		✓			
Admin Sites		✓			
Galinos	✓			✓	
Eopyy	✓	✓		✓	
e-Prescription	✓				
VPN	✓				✓

Πίνακας 3.2: Υπηρεσίες τις οποίες χρησιμοποιούν οι χρήστες κάθε κατηγορίας

Μια σύντομη επεξήγηση των υπηρεσιών που φαίνονται στον πίνακα 3.2:

- **HIS:** Το HIS (Hospital Information System) πρόκειται για το βασικό νοσοκομειακό πληροφοριακό σύστημα και έχει μια ιδιωτική IP η οποία είναι προσβάσιμη μόνο από το εσωτερικό δίκτυο του νοσοκομείου. Τα 2 κυριότερα ports του είναι το 7778 και το 51001. Το port 7778 χρησιμοποιείται κυρίως από διοικητικό προσωπικό, όπως το λογιστήριο, το τμήμα ιατρικών προμηθειών, το τμήμα μισθοδοσίας κ.α., ενώ το port 51001 χρησιμοποιείται από τα τμήματα των διαφόρων κλινικών, όπου οι χρήστες είναι συνήθως γιατροί, νοσοκόμοι ή φαρμακοποιοί.

- **Προμηθεύς:** Ο Προμηθεύς ή promitheus.gov ή eprocurement.gov είναι η κύρια ιστοσελίδα που χρησιμοποιείται από το νοσοκομειακό προσωπικό για ιατρικές προμήθειες, οπότε είναι προφανές ότι οι χρήστες του ανήκουν στο διοικητικό προσωπικό και ειδικότερα στο τμήμα ιατρικών προμηθειών.

- **Γαληνός:** Ο Γαληνός είναι μια ιστοσελίδα που περιέχει οδηγίες για φαρμακευτικά σκευάσματα, οπότε οι χρήστες που την συμβουλεύονται δεν μπορεί παρά να είναι γιατροί ή φαρμακοποιοί.

- **ΕΟΠΥΥ:** Ο eopyy.gov ή όπως δηλώνουν τα αρχικά του, Εθνικός Οργανισμός Παροχής Υπηρεσιών Υγείας, πρόκειται για μια ιστοσελίδα που προσφέρει πολλές ενέργειες σε χρήστες που ανήκουν τόσο σε ιατρικό όσο και σε διοικητικό προσωπικό.

- **E-prescription:** Η ιστοσελίδα αυτή αφορά την ηλεκτρονική συνταγογράφηση φαρμάκων στους ασθενείς και χρησιμοποιείται αποκλειστικά και μόνο από γιατρούς.

- **VPN Πανεπιστημίου Θεσσαλίας:** Το VPN του Πανεπιστημίου μπορεί να χρησιμοποιηθεί μόνο από γιατρούς (νοσοκομειακούς, πανεπιστημιακούς και φοιτητές) και την γραμματεία των κλινικών.

- **Λοιπές ιστοσελίδες:** Λιγότερο σημαντικές ιστοσελίδες (ως προς την συχνότητα εμφάνισης τους στα δεδομένα) όπως το ΗΔΙΚΑ, το ΕΦΚΑ, το apografi.gov, το ebaby.ypes κ.α. χρησιμοποιούνται αποκλειστικά από διοικητικό προσωπικό.

Χάρη σε αυτές τις υπηρεσίες και τις επεξηγήσεις που δόθηκαν από υπευθύνους του νοσοκομείου (σχετικά με το ποια κατηγορία χρηστών έχει πρόσβαση/μπαίνει σε κάθε υπηρεσία), προέκυψαν οι παρακάτω κλάσεις:

- **Doctor**
- **Central Administration**
- **Clinic Administration**
- **Doctor/Clinic Administration**
- **Nurse**
- **Pharmacist**
- **Secretariat VPN**

Οι κλάσεις Central, Clinic Administration προέκυψαν από διάσπαση του Administration για την καλύτερη περιγραφή ορισμένων χρηστών και η κλάση Doctor/Clinic Administration για χρήστες με περίεργο συνδυασμό επισκέψεων σε υπηρεσίες, όπως εξηγείται και στην παρακάτω λίστα κανόνων που δημιουργήθηκε για την αυτόματη ανάθεση κλάσεων στους χρήστες.

Να τονιστεί ότι σε ορισμένα σημεία των κανόνων έγιναν κάποιες παραδοχές λόγω έλλειψης επαρκών πληροφοριών (σημειώνονται όπου έγιναν), οπότε είναι πιθανό να έχει εισαχθεί κάποια μεροληψία (bias) στα δεδομένα.

1. Κίνηση προς το E-prescription πραγματοποιείται από γιατρούς.
2. Κίνηση προς το promitheus.gov πραγματοποιείται από διοικητικό προσωπικό.
3. Κίνηση **μόνο** προς τον Γαληνό πραγματοποιείται από φαρμακοποιούς.
4. Κίνηση **μόνο** προς Admin Sites πραγματοποιείται από διοικητικό προσωπικό.
5. Κίνηση **μόνο** προς Admin Sites και ΕΟΠΥΥ πραγματοποιείται από διοικητικό προσωπικό.
6. Κίνηση **μόνο** προς VPN πραγματοποιείται από την κατηγορία Secretariat VPN η οποία περιέχει γραμματεία κλινικής, γιατρούς και φοιτητές, αλλά λόγω έλλειψη επιπλέον πληροφοριών αποτελεί ξεχωριστή κλάση.
7. Κίνηση που περιλαμβάνει το VPN μπορεί να έχει πραγματοποιηθεί από χρήστες διαφόρων κλάσεων, οπότε με την βοήθεια του πεδίου SRC_MACHINE και τον όγκο της κίνησης προς τις υπόλοιπες υπηρεσίες πραγματοποιήθηκε χειροκίνητη ανάθεση της κλάσης του χρήστη.
8. Κίνηση **μόνο** προς ένα port του HIS πραγματοποιείται από νοσοκόμους για την περίπτωση του 51001 και από διοικητικό προσωπικό στην περίπτωση του 7778.
9. Αν υπάρχει κίνηση προς και τα δύο ports του HIS, τότε για να αποφασιστεί η τελική κλάση ακολουθείται το εξής μοτίβο:
   (α′) Αν ο αριθμός των flows προς το port 7778 είναι τουλάχιστον διπλάσιος από τον αριθμό των flows προς το port 51001 τότε ανήκει στο διοικητικό προσωπικό.

(β′) Αν ο αριθμός των flows προς το port 51001 είναι τουλάχιστον διπλάσιος από τον αριθμό των flows προς το port 7778 τότε ανήκει στους νοσοκόμους.

10. Για την κίνηση που δεν περιορίζεται μόνο στα ports του HIS, πραγματοποιείται ενδελεχής έλεγχος του όγκου της κίνησης προς κάθε υπηρεσία και του πεδίου SRC_MACHINE ώστε να αποφασιστεί αν θα τοποθετηθεί ο χρήστης στην κλάση Doctor, Administration ή Doctor/Clinic Administration.

11. Κίνηση που εμπεριέχει τον Γαληνό, αλλά όχι τις υπηρεσίες HIS, Προμηθέα, VPN και E-prescription, μελετάται όπως και στον κανόνα 10.

12. Κίνηση που εμπεριέχει το HIS, αλλά όχι τις υπηρεσίες Γαληνού, Προμηθέα, VPN και E-prescription, μελετάται όπως και στους κανόνες 9, 10.

13. Κίνηση που εμπεριέχει το HIS και τον Γαληνό, αλλά όχι τις υπηρεσίες Προμηθέα, VPN και E-prescription, μελετάται όπως και στους κανόνες 9, 10.

14. Κίνηση **μόνο** προς τον ΕΟΠΥΥ μπορεί να έχει πραγματοποιηθεί από χρήστες διαφόρων κλάσεων, οπότε με την βοήθεια του πεδίου SRC_MACHINE πραγματοποιήθηκε χειροκίνητη ανάθεση της κλάσης του χρήστη.

15. Κίνηση **μόνο** προς τον ΕΟΠΥΥ και Admin Sites μπορεί να έχει πραγματοποιηθεί από χρήστες διαφόρων κλάσεων, οπότε με την βοήθεια του πεδίου SRC_MACHINE πραγματοποιήθηκε χειροκίνητη ανάθεση της κλάσης του χρήστη.

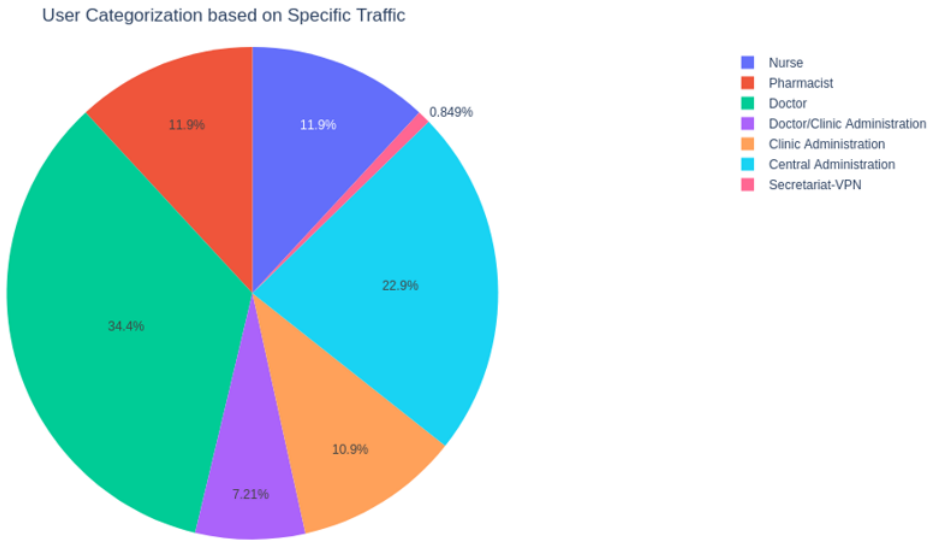
16. Αν ένας χρήστης έχει κατηγοριοποιηθεί στο διοικητικό προσωπικό και ανήκει η συσκευή του σε οποιαδήποτε κλινική, τότε τοποθετείται στην κλάση Clinic Administration.

17. Αν ένας χρήστης έχει κατηγοριοποιηθεί στο διοικητικό προσωπικό και ανήκει η συσκευή του στους τομείς secretariat, medical supplies ή accounting office, τοποθετείται στην κλάση Central Administration.

18. Για χρήστες με περίεργη κίνηση, όπου οι πληροφορίες δεν αρκούσαν για να χαρακτηριστούν γιατροί ή διοικητικό προσωπικό, χρησιμοποιήθηκε η κλάση Doctor/Clinic Administration.

19. Για τους εναπομείναντες 20 χρήστες που δεν υπάγονται σε κανέναν από τους άλλους κανόνες, πραγματοποιήθηκε χειροκίνητη ανάθεση κλάσης με βάση εμπειρικούς κανόνες που απορρέουν από τους παραπάνω.

Αφότου έλαβε μέρος η κατηγοριοποίηση με τους παραπάνω κανόνες, οι 707 χρήστες κατανεμήθηκαν ως εξής στις διάφορες κλάσεις:



Σχήμα 3.16: Διάγραμμα κατανομής χρηστών σε κλάσεις μετά την κατηγοριοποίηση

Κεφάλαιο 4

Συσταδοποίηση

Στο προηγούμενο κεφάλαιο, κατά την επεξεργασία των αρχικών δεδομένων προέκυψε το πεδίο SRC_MACHINE μέσω των DHCP Logs. Για να ελεγχθεί η ορθότητα αυτού του μετασχηματισμού, πραγματοποιήθηκε συσταδοποίηση στην ροή των διαφόρων χρηστών. Δηλαδή αν χρήστες με ίδιο SRC_MACHINE βρεθούν στην ίδια συστάδα, τότε αυτό δείχνει ότι τα DHCP Logs ήταν ακριβή και ότι οι χρήστες αυτοί έχουν παρεμφερή δικτυακή συμπεριφορά.

4.1 Ορισμός Προφίλ Χρήστη για την Συσταδοποίηση

Η έννοια **χρήστης** που προαναφέρθηκε ή αλλιώς **προφίλ χρήστη**, αφορά την ροή ενός ξεχωριστού ID για μια ολόκληρη βάρδια (οι τρεις πιθανές βάρδιες αναφέρθηκαν και στο κεφάλαιο 3), με την υπόθεση ότι η συσκευή που περιγράφεται από το ID δεν χρησιμοποιείται από πολλαπλούς χρήστες. Είναι πιθανό αυτή η υπόθεση να μην ισχύει για όλους τους χρήστες, αφού δεν δουλεύουν όλοι τα ίδια 8ωρα ή ακομή και τον ίδιο αριθμό ωρών, αλλά γίνεται με βάση πληροφορίες από τους υπευθύνους του νοσοκομείου (που επιβεβαιώνουν ότι οι περισσότεροι χρήστες έχουν τις συγκεκριμένες βάρδιες).

4.1.1 Υπηρεσίες προς μελέτη

Επειδή οι υπηρεσίες και οι ιστοσελίδες που χρησιμοποιούνται από τους χρήστες είναι εκατοντάδες και η μελέτη όλων τους είναι πρακτικά αδύνατη, επιλέχθηκαν μερικές από τις πιο συνηθισμένες νοσοκομειακές υπηρεσίες, όπως το HIS, το Πληροφοριακό Σύστημα Εργαστηρίων (Laboratory Information System, LIS), τα Συστήματα Διαχείρισης Κτιρίων (Building Management Systems, BMS) και η Ψηφιακή Απεικόνιση και Επικοινωνιές στην Ιατρική (Digital Imaging and COmmunications in Medicine, DICOM).

Η υπηρεσία HIS εξηγήθηκε και στο προηγούμενο κεφάλαιο, αλλά μια σύντομη περιγραφή των υπολοίπων είναι η εξής:

- Το **LIS** είναι ένα σύστημα λογισμικού που προσφέρει χρήσιμα χαρακτηριστικά για την υποστήριξη των λειτουργιών ενός ιατρικού εργαστηρίου. Μερικά σημαντικά χαρακτηριστικά είναι η διαχείριση δειγμάτων εργαστηρίων, διεπαφές ανταλλαγής δεδομένων και η ενσωμάτωση διαφόρων συσκευών για την ηλεκτρονική μεταφορά τους.

- Το **BMS** είναι ένα σύστημα ελέγχου που ελέγχεται από υπολογιστές και είναι εγκατεστημένο σε κτίρια με σκοπό τον έλεγχο και την παρακολούθηση ηλεκτρομηχανολογικού εξοπλισμού όπως συστήματα ασφαλείας, εξαερισμού, φωτισμού, πυρόσβεσης κ.α.

- **DICOM** ονομάζεται το πρότυπο που χρησιμοποιείται για την επικοινωνία και την διαχείριση της πληροφορίας ιατρικών εικόνων και σχετικών δεδομένων. Χρησιμοποιείται κυρίως για την αποθήκευση και την μεταφορά των ιατρικών εικόνων, επιτρέποντας την ενσωμάτωση συσκευών όπως scanners, εκτυπωτών, servers, workstations κ.α. από διαφορετικούς κατασκευαστές.

4.1.2 Πεδία προς μελέτη

Για την προσωμοίωση της δικτυακής συμπεριφοράς ενός χρήστη χρησιμοποιήθηκαν τρία πεδία (που προέκυψαν με τους κατάλληλους μετασχηματισμούς των δεδομένων μας), τα οποία θεωρήθηκε ότι περιγράφουν αρκετά καλά την δικτυακή συμπεριφορά του χρήστη. Αυτά τα πεδία είναι:

- **Διάρκεια ροής (Flow Duration)**, η οποία προκύπτει από την διαφορά των πεδίων FLOW_END_MILLISECONDS και FLOW_START_MILLISECONDS και μετριέται σε milliseconds.
- **Flow Interarrival Interval** που προκύπτει από την διαφορά του πεδίου FLOW_START_MILLISECONDS μεταξύ δύο διαδοχικών δεδομένων ροής του ίδιου χρήστη.
- **Flow Bytes** που προέκυψαν με άθροισμα των πεδίων IN_BYTES και OUT_BYTES.

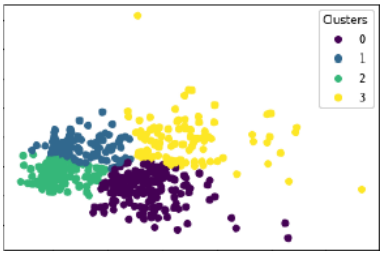
Σχετικές εργασίες που ακολουθούν παρεμφερή μοντελοποίηση με αυτή της παρούσας διπλωματικής είναι οι [16], [17], [18].

4.2 Συσταδοποίηση με αλγορίθμο k-Means και Gaussian Mixture Models

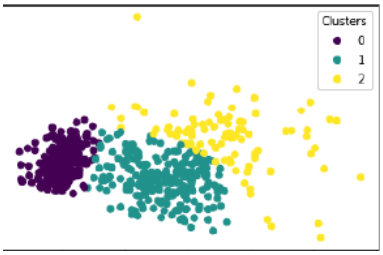
Πριν πραγματοποιηθεί η συσταδοποίηση με k-Means και Gaussian Mixture Models, έπρεπε να αναπαρασταθούν με κάποιο τρόπο οι κατανομές των δεδομένων του κάθε χρήστη (διότι ο κάθε χρήστης είχε περισσότερα από 1 flows). Για να επιτευχθεί αυτό εκπαιδεύτηκαν αρχικά γενικά μοντέλα μείξης (General Mixture Models), από τα οποία προέκυψαν κάποια βάρη για την κατανομή καθενός εκ των τρίων πεδίων (Duration, Interarrival Interval, Bytes) κάθε χρήστη. Επίσης σαν εξτρά δεδομένα δόθηκαν και πληροφορίες όπως η βάρδια, το αν είναι καθημερινή, Σαββατοκύριακο ή/και αργία, με σκοπό να βοηθήσουν στον διαχωρισμό των χρηστών.

Έπειτα με την βοήθεια των συναρτήσεων εύρεσης βέλτιστου αριθμού συστάδων (Elbow Method για k-Means και BIC για Gaussian Mixture Models), βρέθηκε το καλύτερο μοντέλο το οποίο οπτικοποιήθηκε μετά από μείωση διαστατικότητας σε 2 διαστάσεις με Principal Component Analysis και standardization. Παρακάτω παρουσιάζονται τα αποτελέσματα και τα συμπεράσματα για καθεμιά από τις 4 επιλεγμένες υπηρεσίες.

4.2.1 HIS



(α′) k-Means Clustering



(β′) GMM Clustering

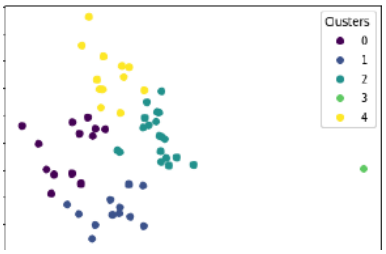
Σχήμα 4.1: Διαγράμματα συσταδοποίησης υπηρεσίας HIS με k-Means και GMM

Στην περίπτωση του k-Means Clustering, 4.1α′, τα δεδομένα διαχωρίστηκαν σε 4 συστάδες. Οι συστάδες 0 και 3 που βρίσκονται στο δεξίο μέρος του διαγράμματος, περιέχουν κυρίως χρήστες από τον τομέα των επειγόντων και δευτερευόντως της γραμματείας. Επίσης σχεδόν όλοι οι χρήστες αυτών των συστάδων χρησιμοποιούν το port 7778, που όπως αναφέρθηκε στο κεφάλαιο 3 χρησιμοποιείται κυρίως από ιατρικό προσωπικό. Οι συστάδες 1, 2 είναι εκ διαμέτρου αντίθετες, με τους κύριους χρήστες να είναι από το λοιμωδών και την παθολογική (και λίγοι από επείγοντα), αλλά με την πλειονότητα τους να χρησιμοποιεί το port 51001.

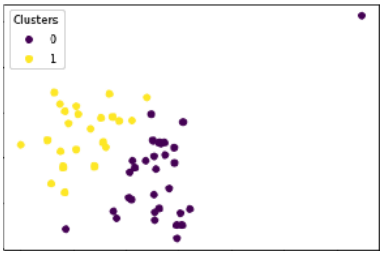
Όμοια, αλλά με μία λιγότερη συστάδα, είναι τα συμπεράσματα από το GMM Clustering, 4.1β′. Σε αυτή την περίπτωση οι συστάδες 1 και 2 περιέχουν τους χρήστες του τομέα των επειγόντων και της γραμματείας (με το 7778 port), ενώ η συστάδα 0 τους χρήστες του λοιμωδών και της παθολογικής κλινικής (με το 51001 port).

Οπότε, εφόσον και οι δύο μέθοδοι συσταδοποίησης συμφωνούν ως προς τον διαχωρισμό των ports 7778 και 51001, επιβεβαιώνονται με ασφάλεια οι πληροφορίες που δόθηκαν από τους υπευθύνους του νοσοκομείου. Επίσης φαίνεται, έστω και ελάχιστα, η υπεροχή του GMM έναντια του k-Means, αφού μπόρεσε να περιγράψει τα ίδια ακριβώς πράγματα με λιγότερες συστάδες.

4.2.2 DICOM



(α′) k-Means Clustering



(β′) GMM Clustering

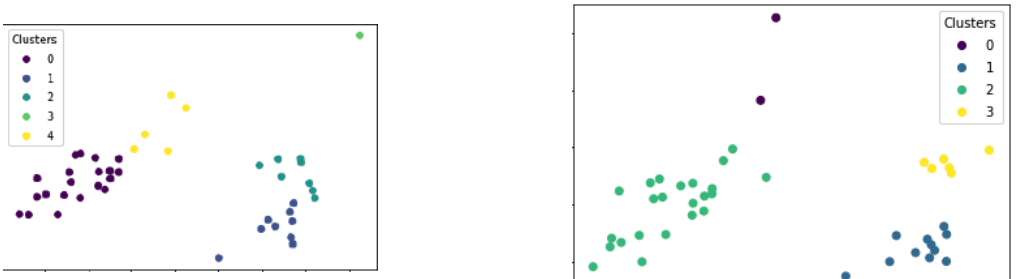
Σχήμα 4.2: Διαγράμματα συσταδοποίησης υπηρεσίας DICOM με k-Means και GMM

Στην υπηρεσία DICOM υπάρχουν 3 τύποι συσκευών, οι xray detectors, οι τομογράφοι και τα διαγνωστικά μηχανήματα. Όπως φαίνεται στο διάγραμμα 4.2α′, η συσταδοποίηση με k-Means δημιούργησε 5 συστάδες, εκ των οποίων οι δύο πρώτες (0, 1) περιέχουν κυρίως τομογράφους και διαγνωστικά, ενώ η 2 και η 4 κυρίως xray detectors και διαγνωστικά. Η 3 περιέχει μόνο μία συσκευή, η οποία θεωρείται outlier.

Στην συσταδοποίηση με GMM, 4.2β′ οι συστάδες που προκύπτουν είναι μόλις 2 και επιτυγχάνεται πλήρης διαχωρισμός των xray detectors και των τομογράφων, με τα διαγνωστικά μηχανήματα να είναι ισομοιρασμένα στις 2 συστάδες.

Είναι φανερό ότι και σε αυτή την περίπτωση ο αλγόριθμος GMM υπερτερεί του k-Means, αφού επιτυγχάνει παρεμφερή αποτελέσματα με λιγότερες συστάδες. Βέβαια και οι 2 αλγόριθμοι αποτυγχάνουν να διαχωρίσουν τα διαγνωστικά μηχανήματα από τις υπόλοιπε συσκευές, κάτι που υποδεικνύει μη σταθερή συμπεριφορά από την συγκεκριμένη κατηγορία.

4.2.3 LIS



(α′) k-Means Clustering (β′) GMM Clustering

Σχήμα 4.3: Διαγράμματα συσταδοποίησης υπηρεσίας LIS με k-Means και GMM

Στην υπηρεσία LIS, υπάρχουν 3 τύποι συσκευών, οι server, workstation, analyser. Στην συσταδοποίηση με k-Means, ο αριθμός των συστάδων που προέκυψε είναι 5, και επετεύχθει καλός διαχωρισμός των δεδομένων, αφού μόνο στην συστάδα 2 υπήρχαν δεδομένα από διαφορετικούς τύπους συσκευών (analyser, workstation). Στην συσταδοποίηση με GMM, 4.3β′, από τις 4 συστάδες που δημιουργήθηκαν, μόνο η 3 είχε έναν τύπο συσκευής (analyser), ενώ οι άλλες 3 είχαν μια μίξη των άλλων 2 τύπων συσκευών (server, workstation). Οπότε σε αυτή την περίπτωση τα αποτελέσματα ήταν καλύτερα με τον αλγόριθμο k-Means.

4.2.4 BMS

(α′) k-Means Clustering (β′) GMM Clustering

Σχήμα 4.4: Διαγράμματα συσταδοποίησης υπηρεσίας BMS με k-Means και GMM

Στην υπηρεσία BMS υπάρχουν 2 τύποι συσκευών, οι server, monitor workstation. Στην συσταδοποίηση με k-Means, 4.4α′, παρατηρείται ότι η μία συστάδα εμπεριέχει μόνο monitor workstations, αλλά η άλλη έχει εκτός από servers και λίγα monitor workstations. Συγκριτικά, η συσταδοποίηση με GMMs, που φαίνεται και στο διάγραμμα 4.4β′, επιτυγχάνει πλήρη διαχωρισμό των δύο κατηγοριών.

4.2.5 Συμπεράσματα

Με αυτή την προσέγγιση παρατηρήθηκε καλός διαχωρισμός των διαφόρων χρηστών σε συστάδων, με φανερή υπεροχή της μεθόδου με GMMs. Οι υπηρεσίες όπου παρατηρήθηκε κάποια δυσκολία στον διαχωρισμό των χρηστών ήταν η LIS και η DICOM, λόγω των συσκευών workstation, diagnostic workstation αντίστοιχα. Παρ΄ όλα αυτά η προσέγγιση αυτή είναι αρκετά γρήγορη, κάτι που αποτελεί μεγάλο πλεονέκτημα συγκριτικά με την παρακάτω, ειδικά σε υπηρεσίες με πολλά δεδομένα.

4.3 Συσσωρευτική συσταδοποίηση με Wasserstein Distance

Μια άλλη προσέγγιση που δοκιμάστηκε ήταν η εύρεση της απόστασης μεταξύ των σημείων των χρηστών. Επειδή οι γνωστές αποστάσεις/μετρικές όπως η Ευκλείδια, η Malahanobis και άλλες αφορούν τις αποστάσεις μεταξύ μοναδικών σημείων (και όπως προαναφέρθηκε ο κάθε χρήστης έχει περισσότερα από ένα σημεία) και η απόκλιση Kullback-Leibler, παρ΄ όλο που μπορεί να βρει αποστάση μεταξύ κατανομών, δεν είναι μετρική, επιλέχθηκε η απόσταση Wasserstein, η οποία εξηγήθηκε και στο κεφάλαιο 2.

Για να μπορέσει όμως να χρησιμοποιηθεί η απόσταση Wasserstein στην συσταδοποίηση, επιλέχθηκε η συσσωρευτική συσταδοποίηση της οποίας ο αλγόριθμος λειτουργεί με αρκετούς τύπους αποστάσεων, ακόμη και ασυνήθιστες όπως η Wasserstein (ο αλγόριθμος k-Means δεν δουλεύει σωστά χωρίς την Ευκλείδια απόσταση γι΄ αυτό και απορρίφθηκε).

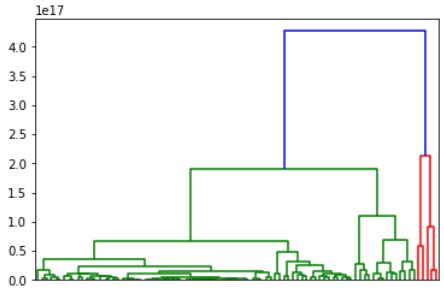
Στην αρχή υπολογίστηκε η απόσταση μεταξύ των κατανομών όλων των δυάδων χρηστών (θεωρώντας ότι κάθε σημείο των δεδομένων του χρήστη είναι ισοπίθανο, εφόσον δεν έχουμε στοιχείο για το αντίθετο). Στην συνέχεια με χρήση δενδρογραμμάτων βρέθηκε ο βέλτιστος αριθμός συστάδων και στο τέλος έγινε οπτικοποίηση των αποτελεσμάτων με χρήση χάρτη θερμότητας (heatmap).

Παρακάτω παρουσιάζονται τα αποτελέσματα και τα συμπεράσματα για 3 από τις 4 επιλεγμένες υπηρεσίες (το HIS αποκλείστηκε ελλείψει επαρκών υπολογιστικών πόρων και θα εξηγηθεί περαιτέρω στα συμπεράσματα η πολυπλοκότητα της μεθόδου με Wasserstein Distance).

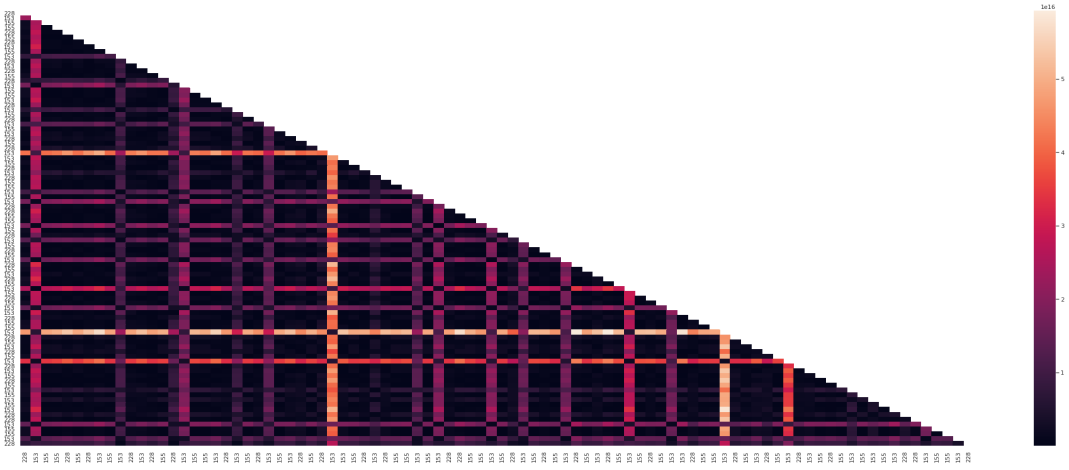
4.3.1 DICOM

Με την βοήθεια του δενδρογράμματος 4.5, αποφασίζεται ο βέλτιστος αριθμός συστάδων, που στην προκειμένη περίπτωση είναι 4.

Αναλύωντας τους χρήστες/συσκευές κάθε συστάδας και το διάγραμμα θερμότητας 4.6, συμπεραίνεται ότι η πλειονότητα των διαγνωστικών μηχανημάτων (diagnostic workstations) διαφέρει αρκετά από τους τομογράφους και τους xray detectors. Αυτό προκύπτει από το γεγονός ότι τα διαγνωστικά μηχανήματα εμφανίζονται και στις 4 συστάδες, ενώ οι xray detectors και οι τομογράφοι μόνο στην 1η, αλλά και από το γεγονός ότι η απόσταση Wasserstein μεταξύ των διαγνωστικών μηχανημάτων με τις άλλες συσκευές είναι μεγαλύτερη και έτσι επηρεάζεται η συσταδοποίηση.



Σχήμα 4.5: Δενδρόγραμμα χρηστών για την υπηρεσία DICOM

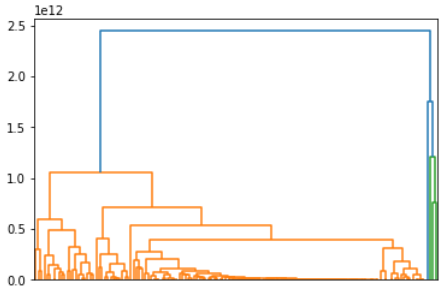


Σχήμα 4.6: Heatmap Wasserstein απόστασης χρηστών για την υπηρεσία DICOM

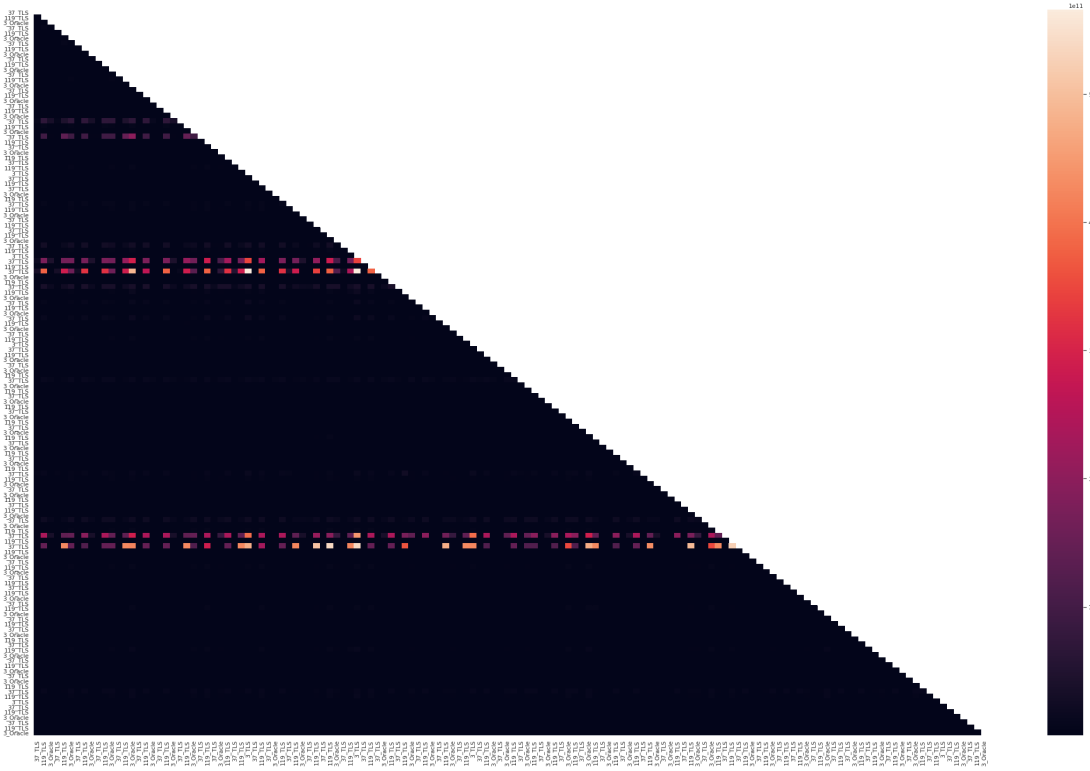
4.3.2 LIS

Με το δενδρόγραμμα 4.7 στους χρήστες του LIS, προκύπτει ότι ο βέλτιστος αριθμός συστάδων είναι 3.

Αναλύωντας και για αυτή την υπηρεσία τους χρήστες/συσκευές κάθε συστάδας και το διάγραμμα θερμότητας 4.8, προκύπτει ότι πλην της 3ης συστάδας που εμπεριέχει μόνο workstations, οι υπόλοιπες 2 είναι μοιρασμένες. Στην 1η περιέχονται servers και analysers, ενώ στην 2η workstations και analysers. Η ομαδοποίηση αυτή δεν είναι κακή αν αναλογιστεί κανείς ότι σε καμιά συστάδα δεν περιέχονται χρήστες και από τις 3 κατηγορίες του SRC_MACHINE, αλλά σίγουρα θα μπορούσε να είναι και καλύτερη.



Σχήμα 4.7: Δενδρόγραμμα χρηστών για την υπηρεσία LIS

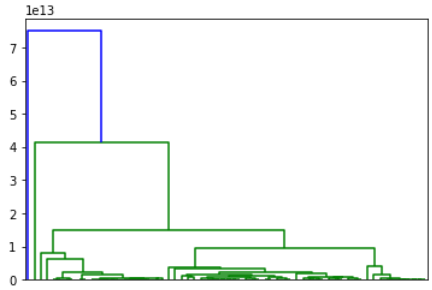


Σχήμα 4.8: Heatmap Wasserstein απόστασης χρηστών για την υπηρεσία LIS

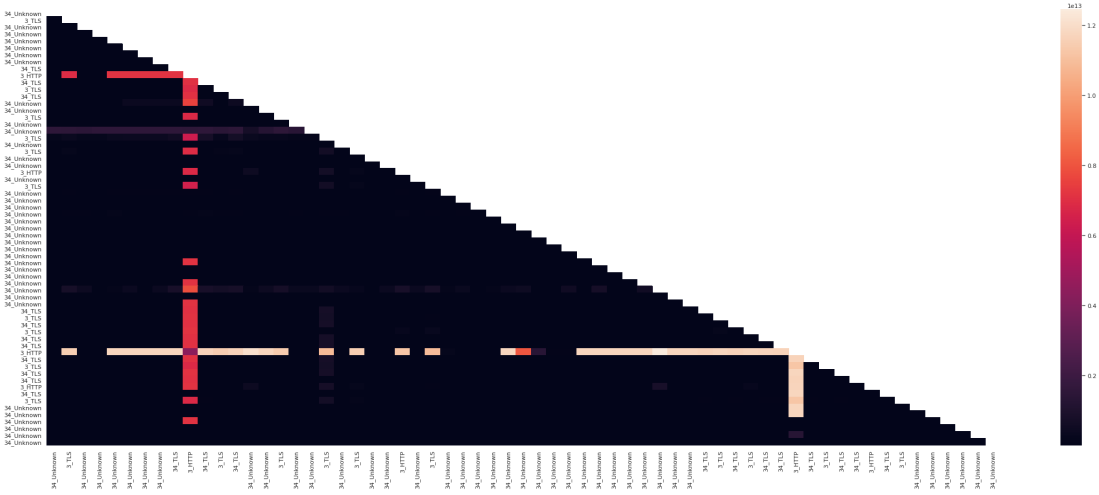
4.3.3 BMS

Με το δενδρόγραμμα 4.9 για τους χρήστες της υπηρεσίας BMS, αποφασίζεται ότι ο βέλτιστος αριθμός συστάδων είναι 3.

Οι 2 εκ των 3 συστάδων αποτελούνται από 1 μόλις χρήστη, με τιμή server στο πεδίο SRC_MACHINE. Η άλλη συστάδα περιέχει τους υπόλοιπους χρήστες (περίπου 50) που ανήκουν στις κατηγορίες server και monitor workstation. Από το διάγραμμα θερμότητας 4.10, φαίνεται ξεκάθαρα ότι ο λόγος για αυτή την ομαδοποίηση είναι η μεγάλη απόσταση των 2 αυτών servers με όλες τις υπόλοιπες συσκευές (και μεταξύ τους), ενώ οι υπόλοιπες συσκευές έχουν μικρή απόσταση μεταξύ τους (αναλογικά με αυτή που έχουν με τους servers). Αυτό εξηγεί την περίεργη αυτή ομαδοποίηση.



Σχήμα 4.9: Δενδρόγραμμα χρηστών για την υπηρεσία BMS



Σχήμα 4.10: Heatmap Wasserstein απόστασης χρηστών για την υπηρεσία BMS

4.3.4 Συμπεράσματα

Αυτό που συμπεραίνεται από αυτή την προσέγγιση είναι ότι αρκετά καλή στην εύρεση διαφορών μεταξύ χρηστών/συσκευών, αλλά όχι τόσο καλή στην εύρεση ομοιοτήτων. Αυτό σημαίνει ότι αν υπάρχουν χρήστες με ακραίες τιμές, θα τους διαχωρίσει από το σύνολο, αλλά σαν αποτέλεσμα αυτού του διαχωρισμού οι υπόλοιποι ομαδοποιούνται στις εναπομείνασες συστάδες (οι οποίες αν είναι λίγες δεν μπορούν να τους διαχωρίσουν καλά).

Το μεγάλο όμως πρόβλημα με αυτή την προσέγγιση είναι η πολυπλοκότητα του υπολογισμού της. Η απόσταση Wasserstein για να υπολογιστεί, έχει πολυπλοκότητα $\mathcal{O}(n^3)$ (όπου n ο αριθμός των flows του χρήστη με τα περισσότερα flows) και ο αριθμός των χρηστών προσθέτει μια επιπλέον πολυπλοκότητα $\mathcal{O}(k^2)$ (όπου k ο αριθμός των χρηστών), για συνολική πολυπλοκότητα $\mathcal{O}(k^2 * n^3)$. Αυτή η πολυπλοκότητα, για υπηρεσίες με πολλούς χρήστες και πολλά flows ανά χρήστη, όπως του HIS, καθιστά την μέθοδο αυτή χρονοβόρα ή/και ανέφικτη.

Κεφάλαιο 5

Παραγωγή Δεδομένων

5.1 Ορισμός Προφίλ Χρήστη για την Παραγωγή Δεδομένων

Όπως και στο προηγούμενο κεφάλαιο, είναι επιτακτική η ανάγκη ορισμού ενός προφίλ χρήστη πάνω στο οποίο θα μοντελοποιηθεί η παραγωγή των δεδομένων. Η βασική διαφορά σε σχέση με το προηγούμενο κεφάλαιο είναι ότι σε αυτό το προφίλ δεν ανταποκρίνεται μόνο σε ένα ID, αλλά σε όλα τα ID's που ανήκουν σε μια ομάδα. Η ομάδα αυτή εξαρτάται από τα πεδία Service (που είναι η υπηρεσία που κάνουν access οι χρήστες), Category (που είναι η κατηγορία που δόθηκε στους χρήστες στο κεφάλαιο 3), Source Machine (που είναι ο τομέας των χρηστών) και Shift (που είναι η βάρδια των χρηστών). Τα δύο τελευταία πεδία μπορούν βέβαια να παραληφθούν σε μελλοντικές επεκτάσεις της διπλωματικής (θα αναφερθεί περαιτέρω στο κεφάλαιο 6) ανάλογα με τις υπολογιστικές δυνατότητες όποιου το επιχειρήσει (διότι ο όγκος των δεδομένων είναι τεράστιος στο στάδιο της εκπαίδευσης).

5.1.1 Υπηρεσίες προς μελέτη

Στις υπηρεσίες προς μελέτη χρησιμοποιήθηκε ένα υπερσύνολο από αυτές του προηγούμενου κεφαλαίου, με την προσθήκη των promitheus.gov, eopyy.gov και Γαληνού, οι οποίες έχουν εξηγηθεί στο κεφάλαιο 3. Ο λόγος που επιλέχθηκαν αυτές είναι γιατί αφορούν νοσοκομειακή κίνηση και είναι πιθανό να προκύψουν χρήσιμα δεδομένα για χρήση σε προσομοιώσεις συμπεριφοράς χρήστη.

5.1.2 Πεδία προς μελέτη

Τα πεδία των δεδομένων που θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων είναι σε πρώτη φάση αυτά που αναφέρθηκαν στο προηγούμενο κεφάλαιο με την διαφορά ότι δεν γίνεται ενοποίηση των bytes για μεγαλύτερη ακρίβεια (Flow Duration, Flow Interarrival Interval, Flow Incoming Bytes και Flow Outgoing Bytes). Σε δεύτερη φάση θα αξιοποιηθεί το conditionality ορισμένων γεννητικών μοντέλων οπότε θα προστεθούν και τα πεδία που ορίζουν την ομάδα ενός χρήστη (Service, Category, Source Machine, Shift).

5.2 Εκπαίδευση μοντέλων σε προφίλ που περιέχουν μόνο αριθμητικά δεδομένα

Τα προφίλ στα οποία αποφασίστηκε να εκπαιδευτούν, ξεχωριστά στο καθένα, τα τρία μοντέλα (GMM, CTGAN, TVAE) φαίνονται στον πίνακα 5.1. Τα προφίλ αυτά επιλέχθηκαν τυχαία ώστε να εμπεριέχεται επαρκής κίνηση από τις βασικές νοσοκομειακές υπηρεσίες και να είναι αντιπροσωπευτικό δείγμα του συνόλου των δεδομένων. Τα γεννητικά μοντέλα βαθιάς μάθησης εκπαιδεύτηκαν όπως ακριβώς περιγράφεται και στο [10], αλλά για 150 αντί για 300 εποχές (διότι παρατηρήθηκε ότι δεν βελτιώνονται τα μοντέλα για περισσότερες εποχές). Για το GMM, αφού βρέθηκε ο βέλτιστος αριθμός components με την μέθοδο BIC, το μοντέλο εκπαιδεύτηκε στα δεδομένα μέχρι την σύγκλιση.

Profile Number	Service	Category	Source Machine	Shift
Profile 1	HIS	Doctor	Pathological	1
Profile 2	HIS	Central Administration	Secretariat	2
Profile 3	Promitheus	Central Administration	Medical Supplies	All
Profile 4	LIS	General	Workstation	1
Profile 5	BMS	General	Monitor Workstation	3
Profile 6	DICOM	General	Diagnostic Workstation	All

Πίνακας 5.1: Προφίλ στα οποία θα εκπαιδευτούν, ξεχωριστά στο καθένα, τα μοντέλα

Όπου οι αριθμοί στα Shifts προκύπτουν από:

- 00:00-08:00 → 1
- 08:00-16:00 → 2
- 16:00-00:00 → 3

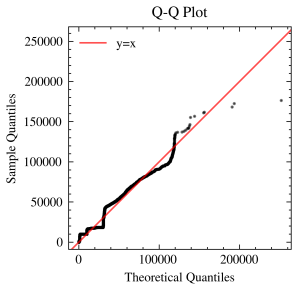
Η κλάση General στα LIS, BMS, DICOM δημιουργήθηκε διότι δεν ήταν γνωστή η κατηγορία του χρήστη αυτών των υπηρεσιών (μπορεί να ήταν οποιαδήποτε από τις υπόλοιπες ή και κάποια που δεν είναι γνωστή).

5.2.1 Αξιολόγηση με Quantile-Quantile διαγράμματα (ανά πεδίο)

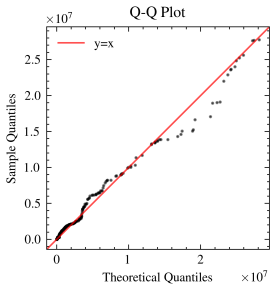
Ένας τρόπος αξιολόγησης των μοντέλων είναι να οπτικοποιήσουμε τα παραγόμενα δεδομένα συγκριτικά με τα πραγματικά. Για την οπτικοποίηση αυτή επιλέχθηκαν τα Quantile-Quantile Plots. Τα Q-Q Plots, όπως αποκαλούνται για συντομία, πρόκειται για διαγράμματα σύγκρισης κατανομών. Το μόνο μειονέκτημα τους, στην περίπτωση που εξετάζεται στην διπλωματική, είναι ότι μπορούν να συγκρίνουν τις κατανομές πραγματικών και συνθετικών δεδομένων ενός πεδίου και όχι και των τεσσάρων. Για αυτό τον λόγο υπάρχει και άλλη μετρική αξιολόγησης των μοντέλων που αξιολογεί και τα τέσσερα πεδία ταυτόχρονα. Παρ΄ όλα αυτά τα Q-Q Plots είναι μια καλή μέθοδος για να παρατηρηθεί κατά πόσο τα συνθετικά δεδομένα άκολουθούν΄ τα πραγματικά, έστω ανά πεδίο.

5.2.1.1 GMM

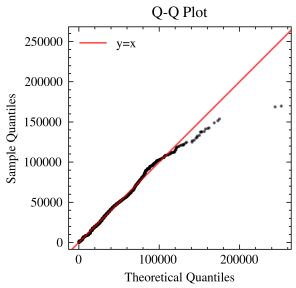
Τα Q-Q Plots των 6 προφίλ που εκπαιδεύτηκαν με GMMs παρουσιάζονται στα διαγράμματα 5.1-5.6.



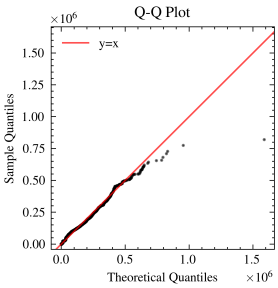
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

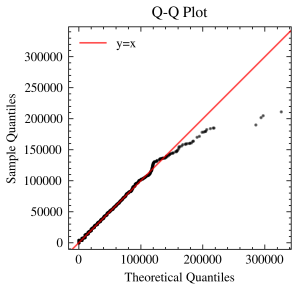


(γ′) Incoming bytes Q-Q Plot

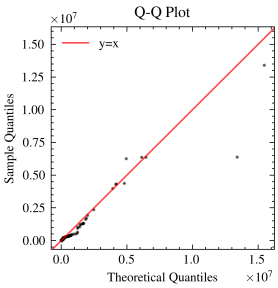


(δ′) Outgoing bytes Q-Q Plot

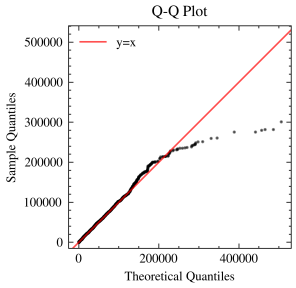
Σχήμα 5.1: Profile 1 Q-Q Plots με GMM



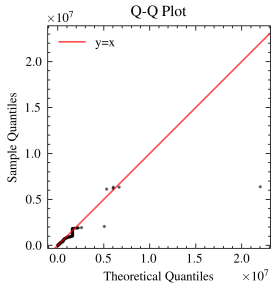
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

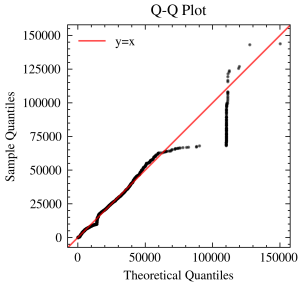


(γ′) Incoming bytes Q-Q Plot

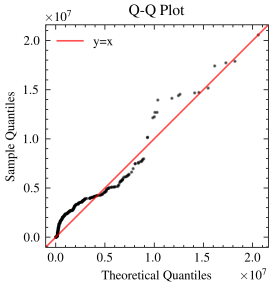


(δ′) Outgoing bytes Q-Q Plot

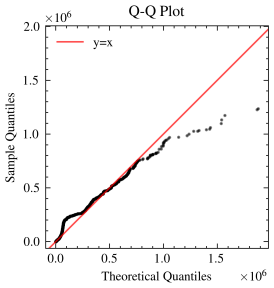
Σχήμα 5.2: Profile 2 Q-Q Plots με GMM



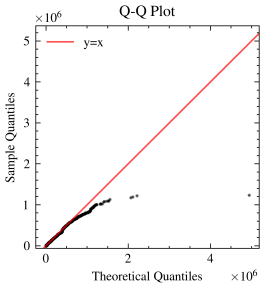
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

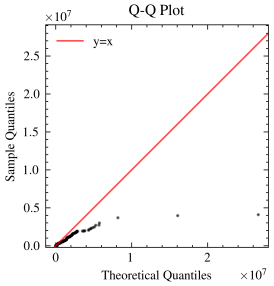


(γ′) Incoming bytes Q-Q Plot

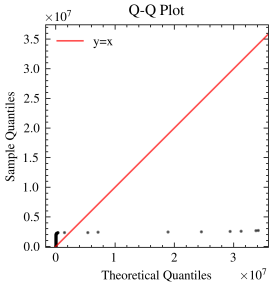


(δ′) Outgoing bytes Q-Q Plot

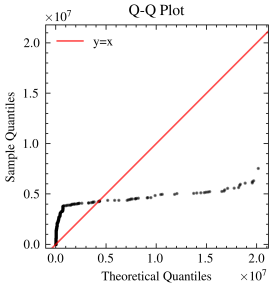
Σχήμα 5.3: Profile 3 Q-Q Plots με GMM



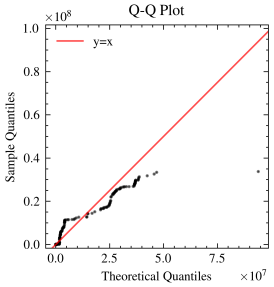
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

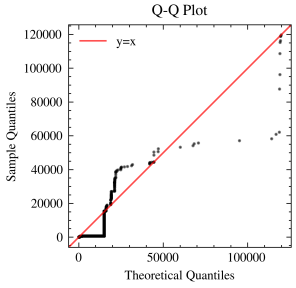


(γ′) Incoming bytes Q-Q Plot

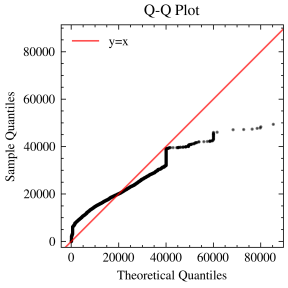


(δ′) Outgoing bytes Q-Q Plot

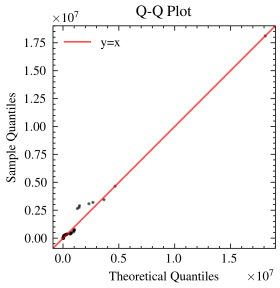
Σχήμα 5.4: Profile 4 Q-Q Plots με GMM



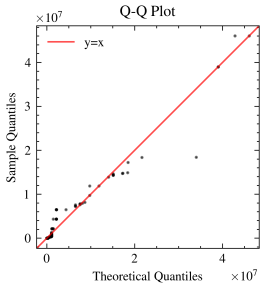
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

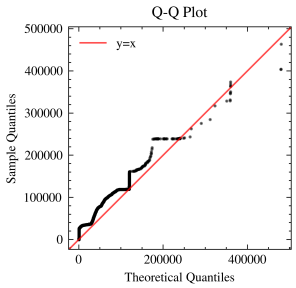


(γ′) Incoming bytes Q-Q Plot

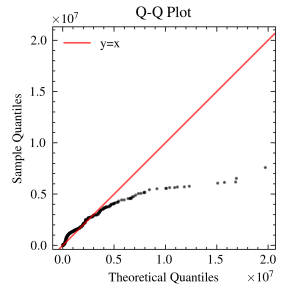


(δ′) Outgoing bytes Q-Q Plot

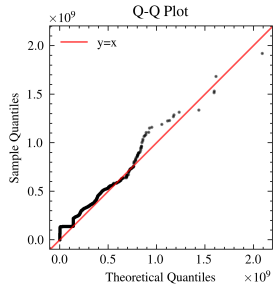
Σχήμα 5.5: Profile 5 Q-Q Plots με GMM



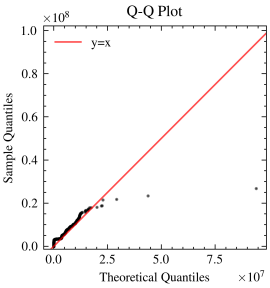
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot



(γ′) Incoming bytes Q-Q Plot

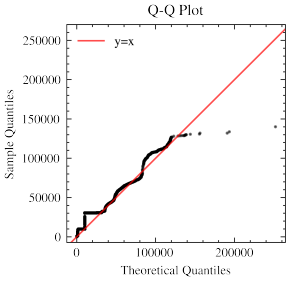


(δ′) Outgoing bytes Q-Q Plot

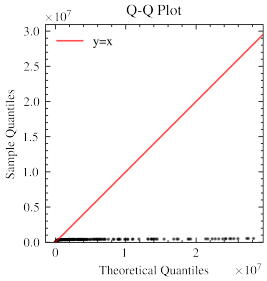
Σχήμα 5.6: Profile 6 Q-Q Plots με GMM

5.2.1.2 CTGAN

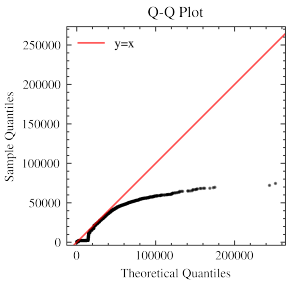
Τα Q-Q Plots των 6 προφίλ που εκπαιδεύτηκαν με CTGANs παρουσιάζονται στα διαγράμματα 5.7-5.12.



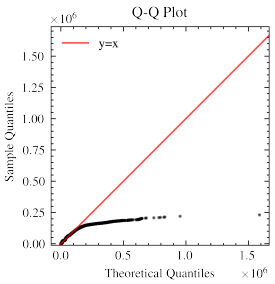
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

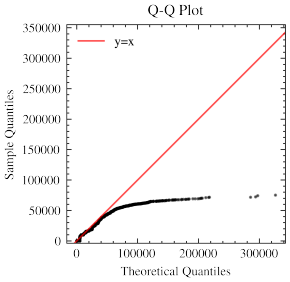


(γ′) Incoming bytes Q-Q Plot

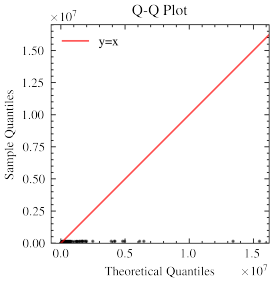


(δ′) Outgoing bytes Q-Q Plot

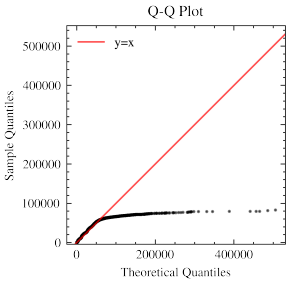
Σχήμα 5.7: Profile 1 Q-Q Plots με CTGAN



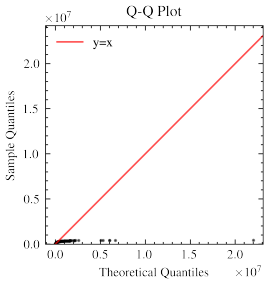
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

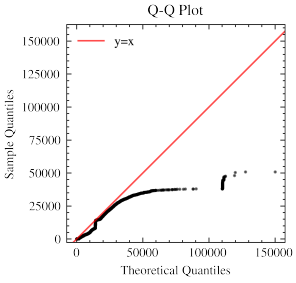


(γ′) Incoming bytes Q-Q Plot

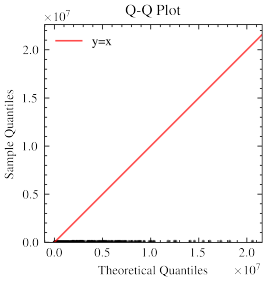


(δ′) Outgoing bytes Q-Q Plot

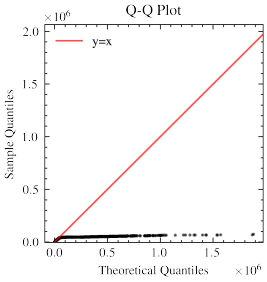
Σχήμα 5.8: Profile 2 Q-Q Plots με CTGAN



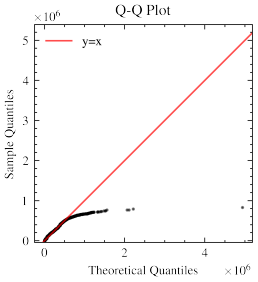
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

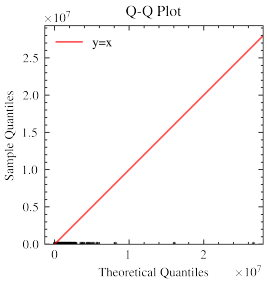


(γ′) Incoming bytes Q-Q Plot

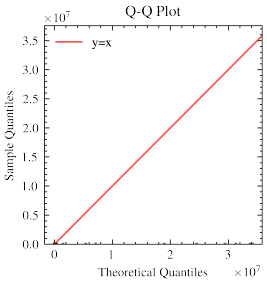


(δ′) Outgoing bytes Q-Q Plot

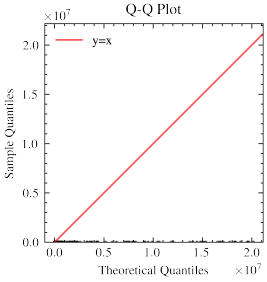
Σχήμα 5.9: Profile 3 Q-Q Plots με CTGAN



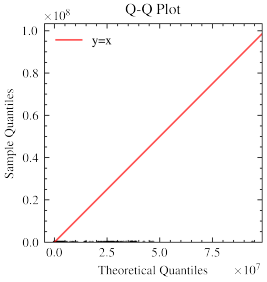
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

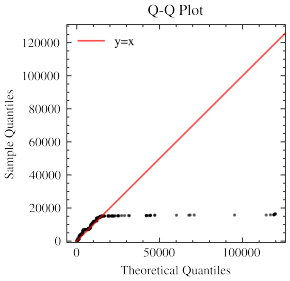


(γ′) Incoming bytes Q-Q Plot

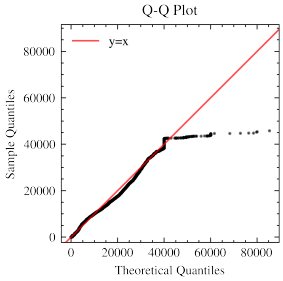


(δ′) Outgoing bytes Q-Q Plot

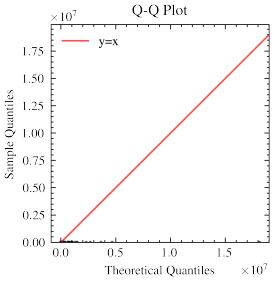
Σχήμα 5.10: Profile 4 Q-Q Plots με CTGAN



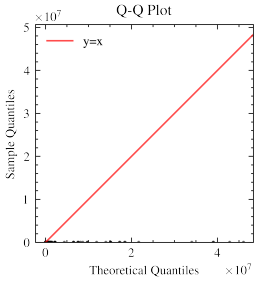
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

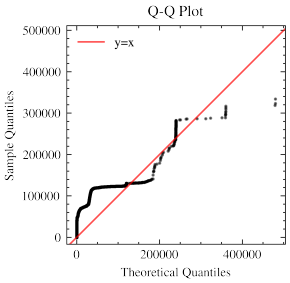


(γ′) Incoming bytes Q-Q Plot

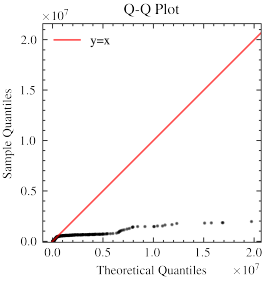


(δ′) Outgoing bytes Q-Q Plot

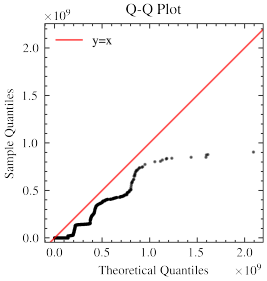
Σχήμα 5.11: Profile 5 Q-Q Plots με CTGAN



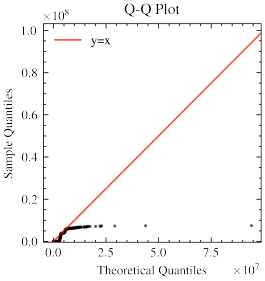
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot



(γ′) Incoming bytes Q-Q Plot

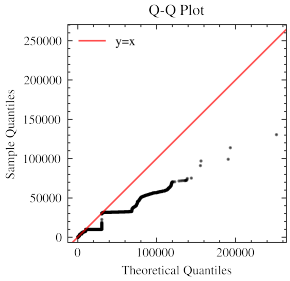


(δ′) Outgoing bytes Q-Q Plot

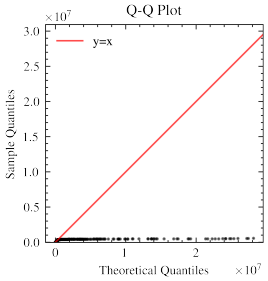
Σχήμα 5.12: Profile 6 Q-Q Plots με CTGAN

5.2.1.3 TVAE

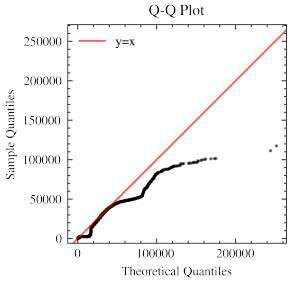
Τα Q-Q Plots των 6 προφίλ που εκπαιδεύτηκαν με TVAEs παρουσιάζονται στα διαγράμματα 5.13-5.18.



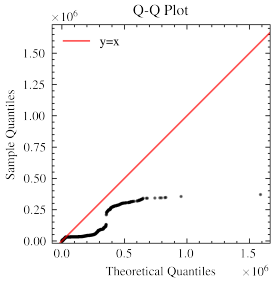
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

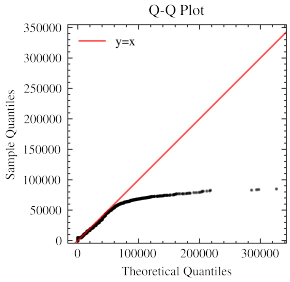


(γ′) Incoming bytes Q-Q Plot

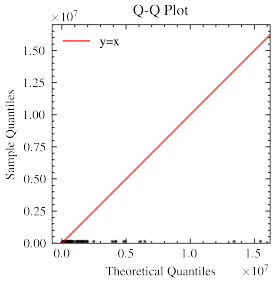


(δ′) Outgoing bytes Q-Q Plot

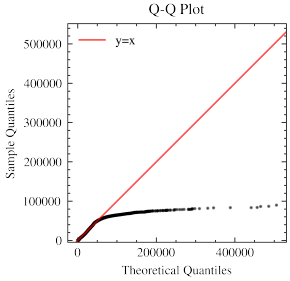
Σχήμα 5.13: Profile 1 Q-Q Plots με TVAE



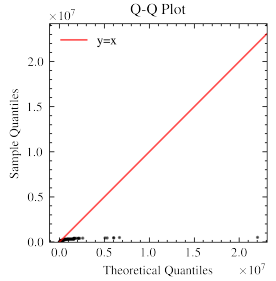
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

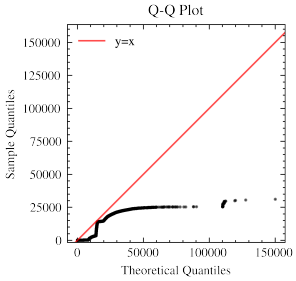


(γ′) Incoming bytes Q-Q Plot

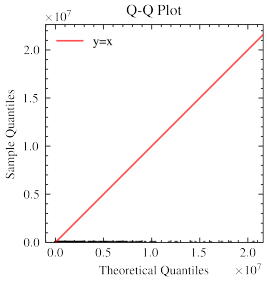


(δ′) Outgoing bytes Q-Q Plot

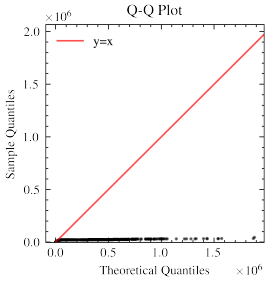
Σχήμα 5.14: Profile 2 Q-Q Plots με TVAE



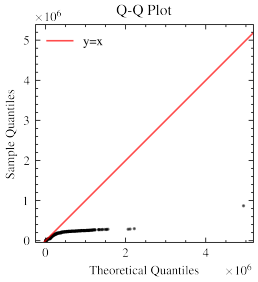
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

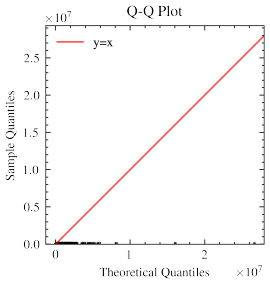


(γ′) Incoming bytes Q-Q Plot

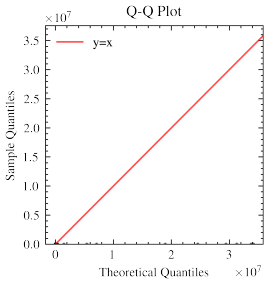


(δ′) Outgoing bytes Q-Q Plot

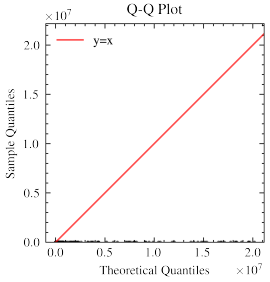
Σχήμα 5.15: Profile 3 Q-Q Plots με TVAE



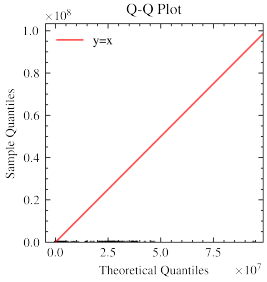
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

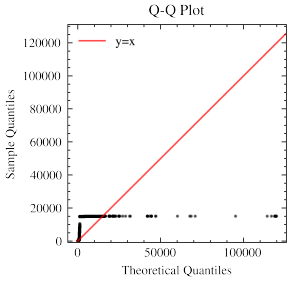


(γ′) Incoming bytes Q-Q Plot

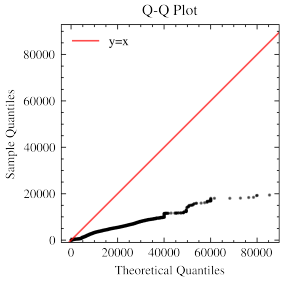


(δ′) Outgoing bytes Q-Q Plot

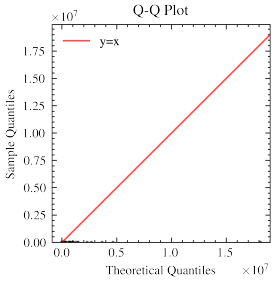
Σχήμα 5.16: Profile 4 Q-Q Plots με TVAE



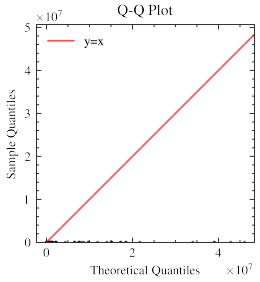
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot

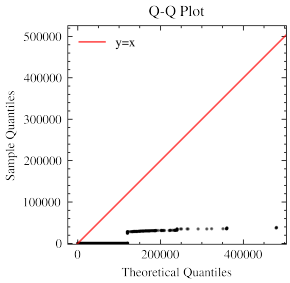


(γ′) Incoming bytes Q-Q Plot

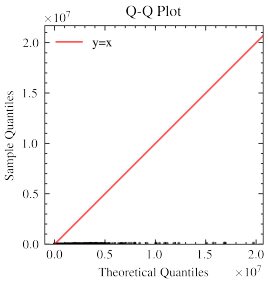


(δ′) Outgoing bytes Q-Q Plot

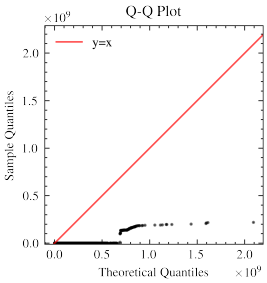
Σχήμα 5.17: Profile 5 Q-Q Plots με TVAE



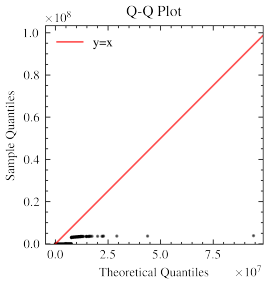
(α′) Duration Q-Q Plot



(β′) Interarrival Interval Q-Q Plot



(γ′) Incoming bytes Q-Q Plot



(δ′) Outgoing bytes Q-Q Plot

Σχήμα 5.18: Profile 6 Q-Q Plots με TVAE

5.2.1.4 Συμπεράσματα

Τα συμπεράσματα που προκύπτουν από τα παραπάνω διαγράμματα είναι τα εξής:

- Είναι φανερό ότι για όλα τα πεδία δεδομένων, οι χαμηλές τιμές απαρτίζουν την πλειονότητα των τιμών. Αυτό είχε δειχθεί και στο κεφάλαιο 3 για τα πεδία Flow Duration, Flow Incoming Bytes, Flow Outgoing Bytes. Για το πεδίο Flow Interarrival Interval που δημιουργήθηκε μετέπειτα, η ανάλυση έδειξε ότι σχεδόν σε όλα τα προφίλ, η πλειονότητα των τιμών ήταν αρκετά κοντά στο 0. Αυτό σημαίνει ότι υπάρχουν αρκετά flows που ξεκινάνε λίγο μετά την εκκίνηση του προηγουμένου από τον ίδιο χρήστη (να σημειωθεί ότι κόπηκαν όλα τα flows από κίνηση τύπου DNS για αυτόν ακριβώς τον λόγο, αλλά και πάλι δεν ήταν αρκετό). Τα παραπάνω γεγονότα οδήγησαν τα μοντέλα, και ιδιαίτερα τα μοντέλα βαθιάς μάθησης, στο να αποκτήσουν bias ως προς τις χαμηλότερες τιμές, εξού και η συγκέντρωση των παραγόμενων δεδομένων σε χαμηλές τιμές ως επί το πλείστον (σε μεγαλύτερο βαθμό από ότι τα πραγματικά δεδομένα).

- Από όλα τα μοντέλα είναι ξεκάθαρο ότι αυτό που προσεγγίζει καλύτερα τις υψηλότερες τιμές είναι το GMM. Βέβαια, επειδή το κάθε πεδίο εξετάζεται ξεχωριστά, δεν γνωρίζουμε αν αυτές οι υψηλές τιμές σε ένα πεδίο, συνδυάζονται με τις κατάλληλες τιμές στα άλλα πεδία (δηλαδή να καθιστούν το flow αντιπροσωπευτικό ενός προφίλ χρήστη).

- Από όλα τα προφίλ αυτό με τα χειρότερα αποτελέσματα στα παραγόμενα δεδομένα (και με τα 3 μοντέλα) είναι της υπηρεσίας LIS (5.4, 5.10, 5.16) ακολουθούμενο από της υπηρεσίας DICOM (5.6, 5.12, 5.18) και BMS (5.5, 5.11, 5.17). Αυτό πιθανώς να οφείλεται εν μέρει στην φύση της κίνησης που προκύπτει από αυτές τις υπηρεσίες, διότι πρόκειται για κίνηση παραγόμενη από ιατρικό εξοπλισμό (στην περίπτωση των LIS, DICOM) και από κτιριακό εξοπλισμό (στην περίπτωση του BMS). Αυτό περιορίζει τις τιμές των πραγματικών δεδομένων, διότι τα μηχανήματα αυτά παράγουν συγκεκριμένες τιμές και καθιστά δύσκολο για τα μοντέλα να εκπαιδευτούν σε αυτές.

- Αντιθέτως στα προφίλ της υπηρεσίας του HIS, παρατηρούνται αρκετά καλύτερα αποτελέσματα, διότι οι τιμές των δεδομένων δεν είναι προκαθορισμένες και είναι περισσότερο διάσπαρτες (δηλαδή ελαττώνεται το bias των μοντέλων προς τις χαμηλότερες τιμές).

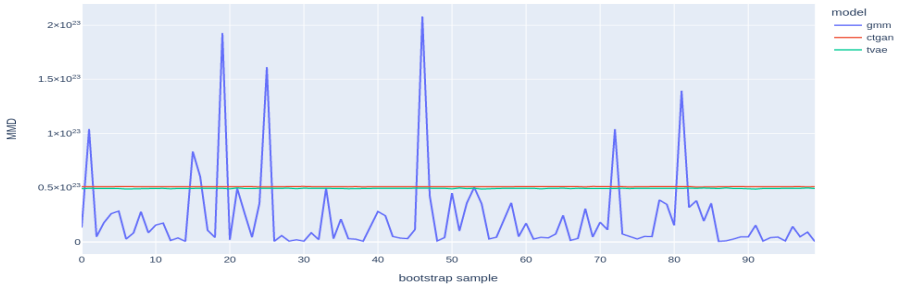
- Όσον αφορά το πεδίο Interarrival Interval, σε αυτό το πεδίο παρατηρούνται τα χειρότερα αποτελέσματα. Επειδή πρόκειται για κατασκευασμένο πεδίο και όχι για δεδομένο που προέκυψε από το εργαλείο nProbe (όπως τα υπόλοιπα), οι τιμές του όπως προαναφέραμε είναι πολύ μικρές. Η πλειονότητα (>80%) των flows που μελετώνται (αυτά με πολλή κίνηση δηλαδή) έχει αμελητέο Interarrival Interval (<1 δευτερολέπτου). Οπότε ειδικά με την αρχιτεκτονική των CTGAN, TVAE, τα οποία εξετάζουν γειτονικά δεδομένα (δοκιμάστηκαν από 500 έως 2000 με παρεμφερή αποτελέσματα), βλέπουν ότι η πλειονότητα αυτών των δεδομένων έχει μικρές τιμές οπότε αποκτούν την προκατάληψη προς τις μικρές τιμές. Βέβαια όπως θα δειχθεί παρακάτω με την άλλη μέθοδο αξιολόγησης, αυτό δεν είναι λάθος υπό το πρίσμα της σύγκρισης και των τεσσάρων πεδίων μαζί.

5.2.2 Αξιολόγηση με MMD (για όλα τα πεδία)

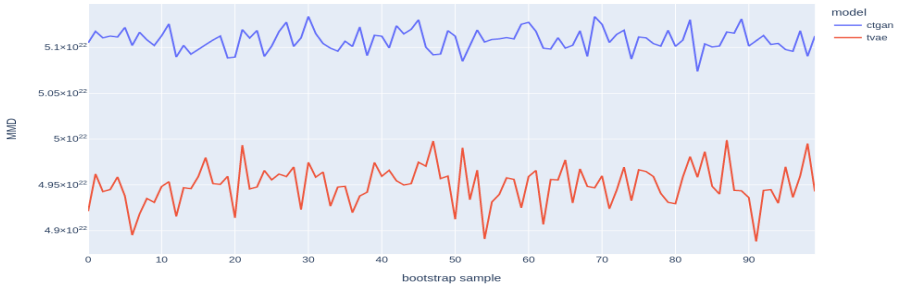
Μια άλλη μέθοδος αξιολόγησης της απόδοσης των μοντέλων είναι αυτή με sampling και MMD. Η λογική πίσω από αυτή την αξιολόγηση είναι να παραχθεί επαρκής αριθμός ανεξάρτητων samples από την κατανομή των συνθετικών δεδομένων και να υπολογιστεί το MMD αυτών των samples από τα πραγματικά δεδομένα. Γίνεται η υπόθεση ότι τα samples (που είναι το 2-5% των συνθετικών δεδομένων, αλλά και των πραγματικών για αυτό το προφίλ χρήστη) θα ακολουθήσουν την κατανομή του συνόλου των συνθετικών δεδομένων για αυτό το προφίλ χρήστη. Θεωρήθηκε ότι 100 samples είναι ικανά για να βγουν συμπεράσματα για την διακύμανση των MMDs κάθε μοντέλου. Σε μερικά από τα προφίλ τα samples εξετάστηκαν σε όλα τα υποσύνολα των πραγματικών δεδομένων (π.χ. στο LIS τα πραγματικά δεδομένα έσπασαν σε 20 υποσύνολα λόγω μεγέθους και για κάθε sample βρέθηκε το MMD από το υποσύνολο) και στο τέλος υπολογίστηκε ο γεωμετρικός μέσος όρος των MMDs για αυτό το sample. Για το MMD χρησιμοποιήθηκε ο πολυωνυμικός πυρήνας (Polynomial kernel) δευτέρου βαθμού.

5.2.2.1 Αποτελέσματα

Τα αποτελέσματα για την εκτέλεση της παραπάνω μεθόδου για κάθε προφίλ παρουσιάζονται στα διαγράμματα 5.19-5.24.

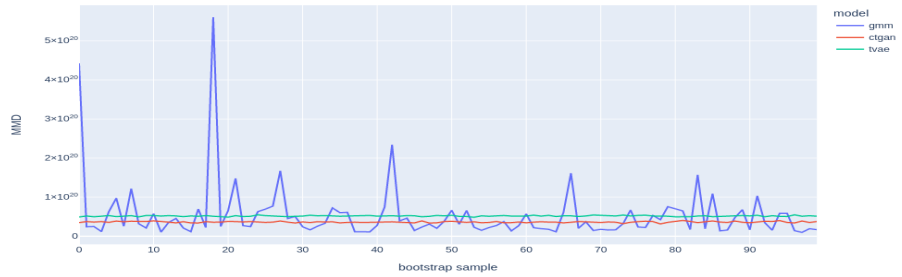


(α′) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων

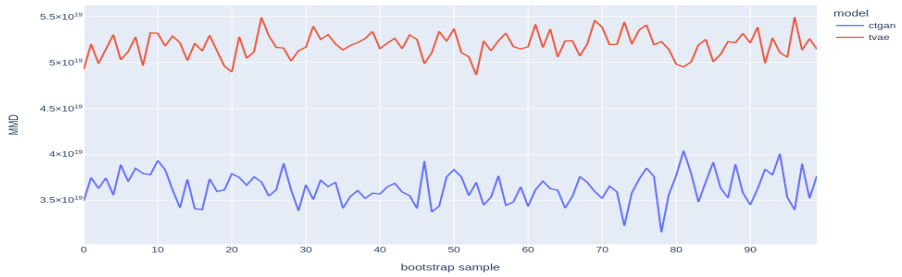


(β′) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.19: Διαγράμματα σύγκρισης των MMD για το προφίλ 1

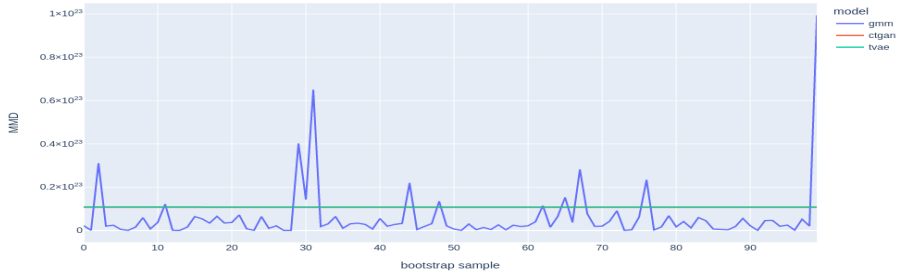


(α΄) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων

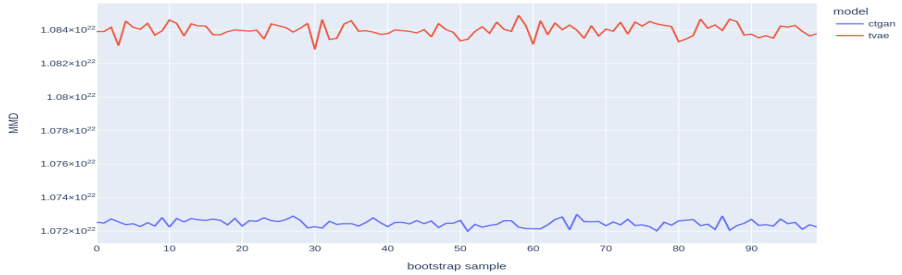


(β΄) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.20: Διαγράμματα σύγκρισης των MMD για το προφίλ 2

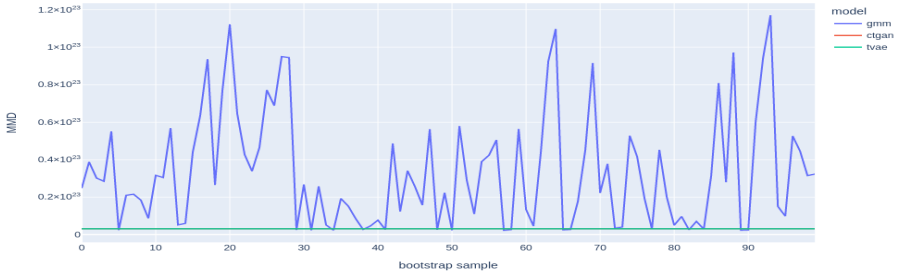


(α΄) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων

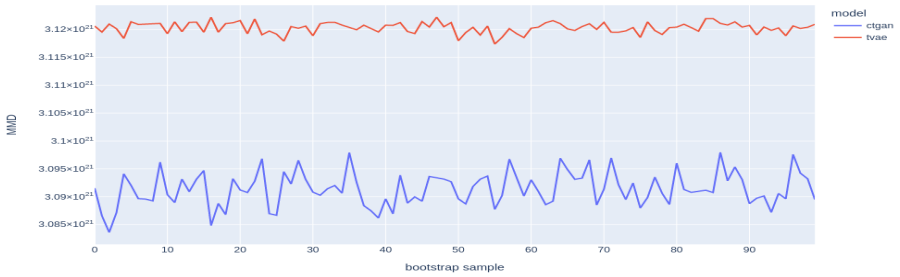


(β΄) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.21: Διαγράμματα σύγκρισης των MMD για το προφίλ 3

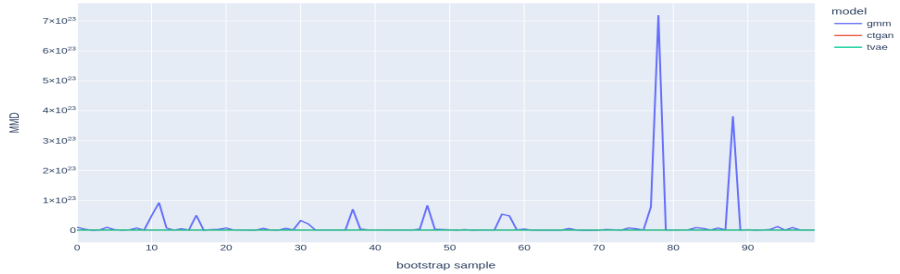


(α′) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων

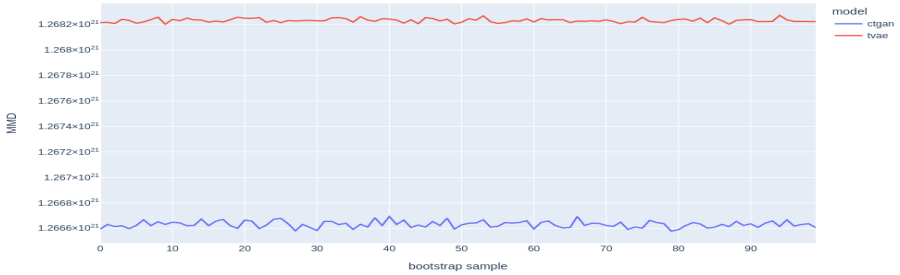


(β′) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.22: Διαγράμματα σύγκρισης των MMD για το προφίλ 4

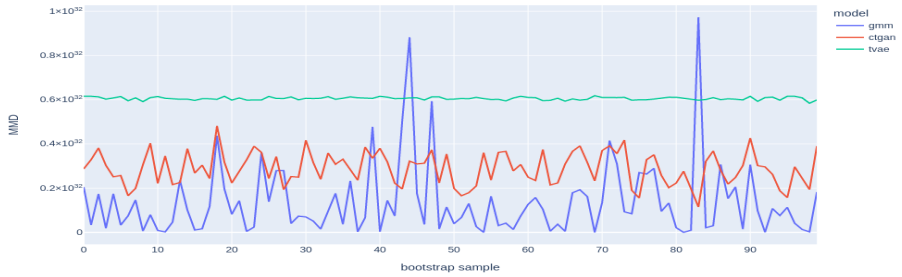


(α′) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων

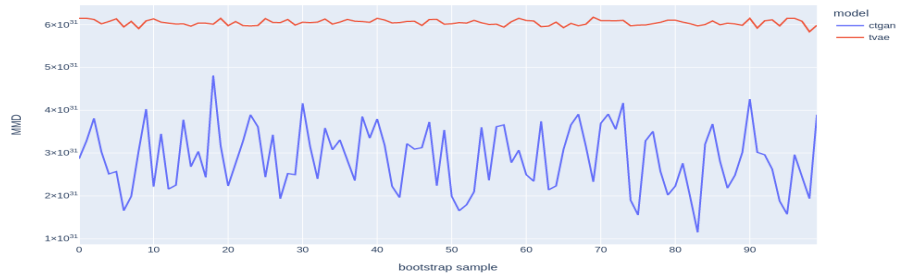


(β′) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.23: Διαγράμματα σύγκρισης των MMD για το προφίλ 5



(α′) Διάγραμμα σύγκρισης των MMD των τριών μοντέλων



(β′) Διάγραμμα σύγκρισης των MMD των CTGAN, TVAE

Σχήμα 5.24: Διαγράμματα σύγκρισης των MMD για το προφίλ 6

Ο λόγος που για κάθε προφίλ παρέχονται δύο διαγράμματα είναι γιατί παρατηρήθηκε αρκετά μικρή διακύμανση της γεωμετρικής μέσης τιμής του MMD μεταξύ των samples των μοντέλων CTGAN, TVAE. Έτσι μπορεί κανείς να διακρίνει πιο εύκολα την διακύμανση αυτών των δύο μοντέλων σε ένα ξεχωριστό διάγραμμα, αλλά και την σύγκριση τους με το GMM στο αρχικό.

5.2.2.2 Συμπεράσματα

Γενικά συμπεράσματα από αυτή την προσέγγιση:

- Με την μετρική MMD μπορεί να γίνει καλύτερη σύγκριση των μοντέλων, αφού συγκρίνονται ταυτόχρονα οι κατανομές και των τεσσάρων πεδίων (Duration, Interarrival Interval, Incoming Bytes, Outgoing Bytes) και όχι του καθενός ξεχωριστά όπως τα διαγράμματα Quantile-Quantile. Η υπεροχή ενός μοντέλου δηλαδή στην σύγκριση ανά πεδίο δεν σημαίνει τίποτα, αφού τα μοντέλα εκπαιδεύονται και στα τέσσερα πεδία μαζί και όχι στο καθένα ξεχωριστά.

- Η χρήση sampling εγγυάται την υπεροχή ενός μοντέλου έναντι ενός άλλου για μεγάλο αριθμό samples (ειδικά όταν δεν υπάρχει μεγάλη διακύμανση μεταξύ των samples).

- Με αυτή την προσέγγιση δεν φαίνεται κάποια υπεροχή από το μοντέλο GMM (ίσα ίσα φαίνεται χειρότερο στην πλειονότητα των προφίλ, με εξαίρεση το 1ο και το 6ο) σε αντίθεση με τα Q-Q Plots που υποδείκνυαν αυτό το πράγμα. Αυτό σημαίνει ότι τα συνθετικά δεδομένα των άλλων δύο μοντέλων παρ′ ότι biased, ακολουθούν την πραγματική κατανομή που είναι συγκεντρωμένη στις χαμηλές τιμές.

- Όσον αφορά τα μοντέλα βαθιάς μάθησης, παρατηρείται μια μικρή υπεροχή του CTGAN έναντι του TVAE σε όλα τα προφίλ εκτός του πρώτου.

- Είναι προφανές από τα διαγράμματα ότι τα μοντέλα CTGAN, TVAE έχουν μικρό variance MMD μεταξύ των samples, κάτι που επιβεβαιώνει την προκατάληψη των μοντέλων στο να παράγουν κοντινές τιμές δεδομένων (που δεν προκαλούν διακύμανση στα MMDs). Αντιθέτως το GMM παράγει πιο μεγάλο εύρος τιμών οπότε είναι πιθανό κάποιες φορές η κατανομή των συνθετικών δεδομένων του να μην ειναι τόσο κοντά στων πραγματικών και γι΄ αυτό υπάρχουν κάποια spikes.

- Το μικρό variance των μοντέλων βαθιάς μάθησης ίσως είναι και αποτέλεσμα overfitting των μοντέλων. Στα GMMs το overfitting αντιμετωπίζεται με το BIC που επιτρέπει την επιλογή του κατάλληλου αριθμού components. Σε αυτά όμως τα μοντέλα, δεν υπάρχει τέτοια δυνατότητα και το γεγονός ότι είναι προσαρμοσμένα για παραγωγή conditional δεδομένων, ίσως τα κάνει χειρότερα σε ένα απλό task παραγωγής unconditional δεδομένων.

Για τους παραπάνω λόγους θα χρησιμοποιηθεί μόνο αυτή η προσέγγιση για την αξιολόγηση των μοντέλων της επόμενης ενότητας, αφού καλύτερα αποτελέσματα στα διαγράμματα Quantile-Quantile δεν συνεπάγονται καλύτερα μοντέλα σύμφωνα με την συγκεκριμένη μετρική.

5.3 Εκπαίδευση conditional μοντέλων σε σύνολο δεδομένων που εμπεριέχει κατηγορικά δεδομένα

Για να μελετηθεί η ικανότητα των μοντέλων CTGAN, TVAE στο να εκπαιδεύονται και να παράγουν δεδομένα conditionally, κάτι που θα ήταν ιδανικό να γίνεται ούτως ή άλλως λόγω της διαδικασίας που απαιτείται για να εκπαιδευτούν τα μοντέλα στο καθένα ξεχωριστά, προστέθηκαν στα υπάρχοντα προφίλ αρκετά ακόμη (ώστε να αυξηθεί το complexity της διαδικασίας εκπαίδευσης και να φανεί πώς ανταποκρίνονται τα μοντέλα) και σε αυτή την προσέγγιση παρέμειναν τα πεδία Service, Category, Source Machine, Shift. Το σύνολο των προφίλ πάνω στα οποία εκπαιδεύτηκαν τα δύο μοντέλα, φαίνονται στον πίνακα 5.2. Τα μοντέλα εκπαιδεύτηκαν για 200 εποχές αντί για 150 που εκπαιδεύτηκαν στην προηγούμενη ενότητα λόγω του μεγαλύτερου όγκου δεδομένων.

Στα προφίλ που αναγράφονται στον πίνακα 5.1 προστέθηκαν και άλλα των ιδίων χρηστών, απλά διαφορετικών υπηρεσιών όπως του Γαληνού, ΕΟΠΥΥ και ακόμη και μη νοσοκομειακών υπηρεσιών όπως το Youtube), αλλά και νέα προφίλ χρηστών και προέκυψε ο πίνακας 5.2. Τα προφίλ όμως που θα αξιολογηθούν στις επόμενες ενότητες θα είναι μόνο αυτά του πίνακα 5.1 (για να υπάρχει μέτρο σύγκρισης). Τα υπόλοιπα προφίλ όπως προαναφέρθηκε, προστέθηκαν ώστε να αυξηθεί η πολυπλοκότητα του προβλήματος που καλούνται τα μοντέλα να λύσουν, δηλαδή την παραγωγή αληθοφανών conditional δεδομένων.

Profile Number	Service	Category	Source Machine	Shift
Profile 1	Ομάδα 2	Doctor	Pathological	All
Profile 2	Ομάδα 4	Central Administration	Secretariat	All
Profile 3	Ομάδα 1	Central Administration	Medical Supplies	All
Profile 4	LIS	General	Workstation	All
Profile 5	BMS	General	Monitor Workstation	All
Profile 6	DICOM	General	Diagnostic Workstation	All
Profile 7	Ομάδα 2	Doctor	Gastroenterology	All
Profile 8	Ομάδα 2	Doctor	Oncology	All
Profile 9	Ομάδα 2	Doctor	Infectious Diseases	All
Profile 10	Ομάδα 4	Clinic Administration	Outpatient Clinics	All
Profile 11	Ομάδα 3	Clinic Administration	Emergency	All
Profile 12	Ομάδα 2	Central Administration	IT	All
Profile 13	Ομάδα 5	Nurse	Cardiology	All
Profile 14	HIS	Nurse	Surgery	All
Profile 15	Ομάδα 1	Central Administration	Accounting Office	All

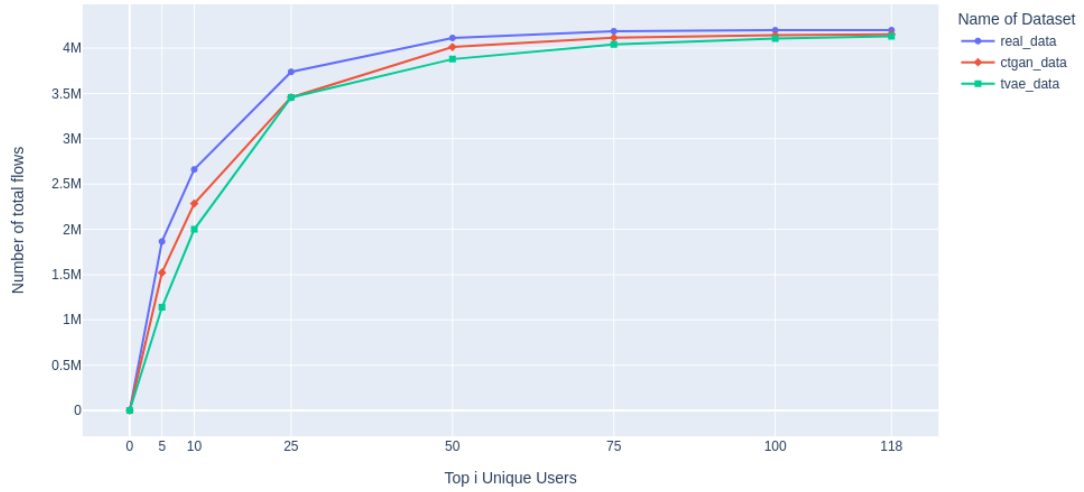
Πίνακας 5.2: Σύνολο δεδομένων από προφίλ στο οποίο θα εκπαιδευτούν τα μοντέλα

Όπου οι ομάδες που αναγράφονται στο service εμπεριέχουν πολλαπλές υπηρεσίες και είναι οι παρακάτω:

- Ομάδα 1: HIS, Promitheus, Youtube
- Ομάδα 2: HIS, Galinos, EOPYY, Youtube
- Ομάδα 3: HIS, Galinos, Youtube
- Ομάδα 4: HIS, EOPYY, Youtube
- Ομάδα 5: HIS, Youtube

5.3.1 Μελέτη παραγόμενων κλάσεων

Ένα πρώτο βήμα για την μελέτη των δύο μοντέλων είναι η ανάλυση των παραγόμενων δεδομένων, όσον αφορά τις κλάσεις που παράγονται. Τα αρχικά δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων είχαν 118 ξεχωριστούς συνδυασμούς χρηστών (ή κλάσεις όπως αναφέρεται παραπάνω). Οπότε σταδιακά, ξεκινώντας από τους 5 κορυφαίους χρήστες σε κάθε σύνολο δεδομένων (πραγματικών και παραγόμενων) και φτάνοντας στους 118, μελετήθηκε το τι ποσοστό των συνολικών flows απαρτίζεται από αυτούς τους i χρήστες στο διάγραμμα 5.25.



Σχήμα 5.25: Συνολικός αριθμός flows των i κορυφαίων (βάσει συχνότητας εμφανίσεων) κλάσεων

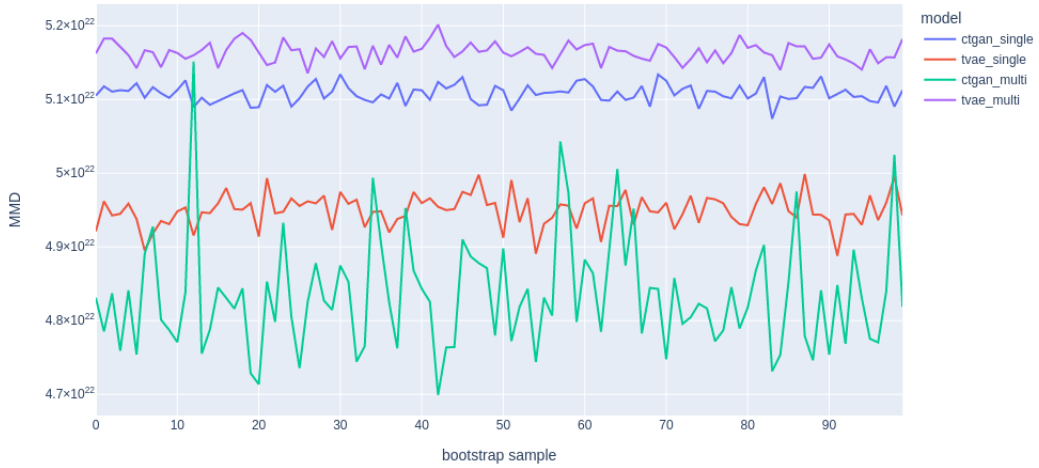
Παρατηρείται ότι το CTGAN ακολουθεί πιο πιστά τις κλάσεις των πραγματικών δεδομένων απ΄ ότι ο TVAE και συνολικά προκύπτει μεγαλύτερο ποσοστό στο τέλος (με μικρή διαφορά βέβαια) ΄πραγματικών΄ κλάσεων. Αυτό δεν είναι βέβαια κάτι καλό ή κακό, διότι έχει να κάνει με την προσέγγιση του κάθε μελετητή. Αν δηλαδή κάποιος χρειαζόταν πιστές κλάσεις, που ανταποκρίνονται περισσότερο στις πραγματικές, τότε το CTGAN θα ήταν το εργαλείο του. Διαφορετικά αν ήθελε μεγαλύτερη ποικιλία στις κλάσεις (αν και κάποιες βέβαια θα έπρεπε να απορριφθούν αν δεν έστεκε καθόλου ο συνδυασμός τους) τότε ο TVAE θα του ήταν πιο χρήσιμος. Παρ΄ όλα αυτά και οι κλάσεις των δεδομένων των δύο μοντέλων δεν αποκλίνουν σημαντικά από των πραγματικών και οι νέες παραγόμενες, αν και πολλές, καλύπτουν ένα πολύ μικρό ποσοστό των δεδομένων (λιγότερο του 3-4%).

5.3.2 Αξιολόγηση με MMD (για όλα τα πεδία)

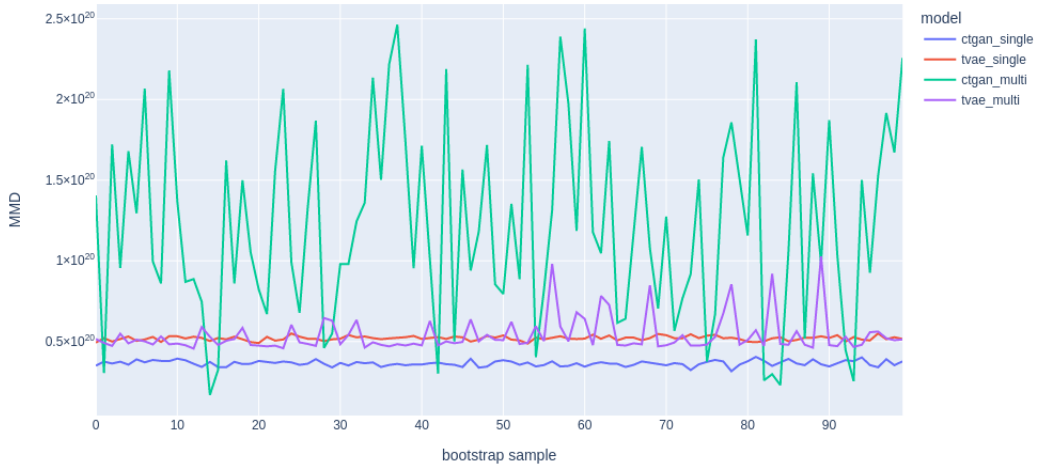
Με την ίδια μετρική αξιολόγησης που χρησιμοποιήθηκε και στην προηγούμενη ενότητα, αξιολογήθηκαν τα 6 αρχικά προφίλ, που όμως εδώ προκύπτουν μετά από την εκπαίδευση σε πολλαπλά προφίλ χρηστών (5.2) και όχι σε κάθε ένα ξεχωριστά.

5.3.2.1 Αποτελέσματα

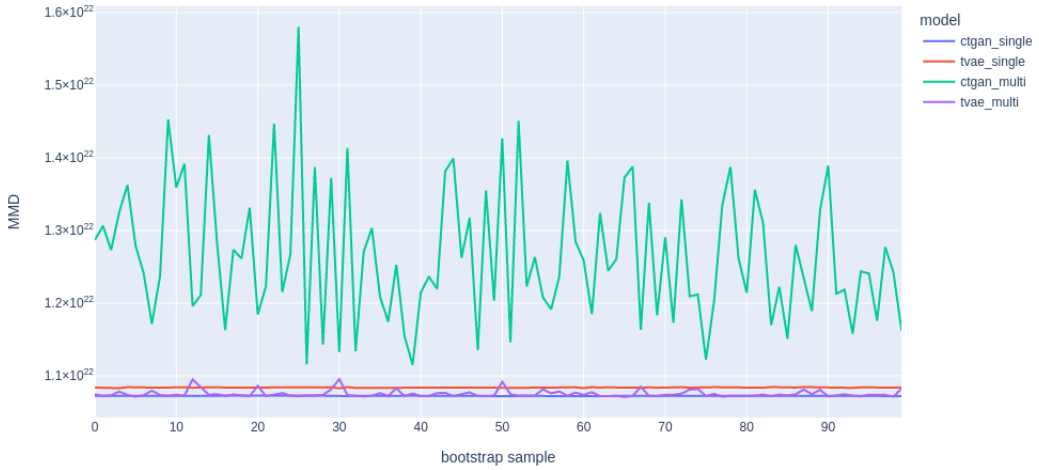
Τα αποτελέσματα για την εκτέλεση της παραπάνω μεθόδου για κάθε προφίλ παρουσιάζονται στα διαγράμματα 5.26-5.31. Τα 4 μοντέλα που συγκρίνονται είναι τα CTGAN, TVAE εκπαιδευμένα στα νέα δεδομένα (με ονόματα ctgan_multi, tvae_multi) και τα CTGAN, TVAE εκπαιδευμένα σε κάθε ένα από τα 6 προφίλ ξεχωριστά (με ονόματα ctgan_single, tvae_single) όπως και είχαν οριστεί στην προηγούμενη ενότητα.



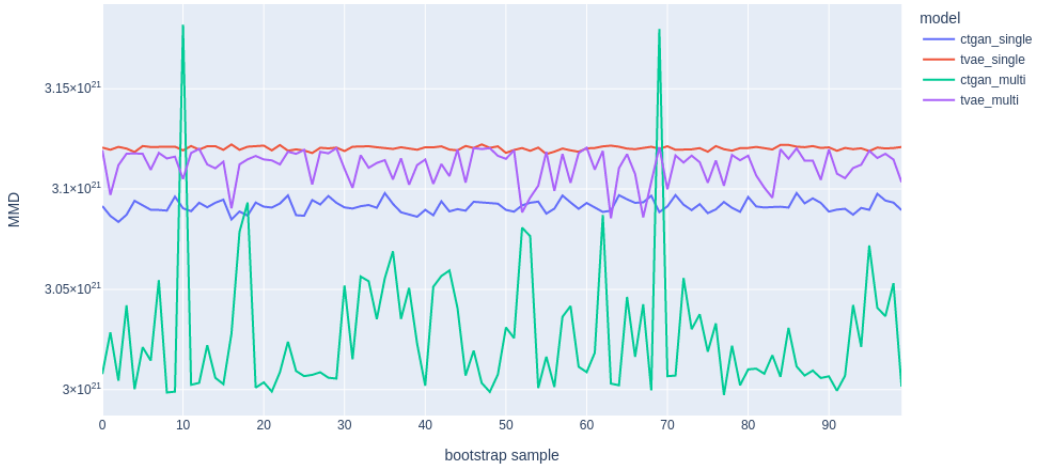
Σχήμα 5.26: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 1



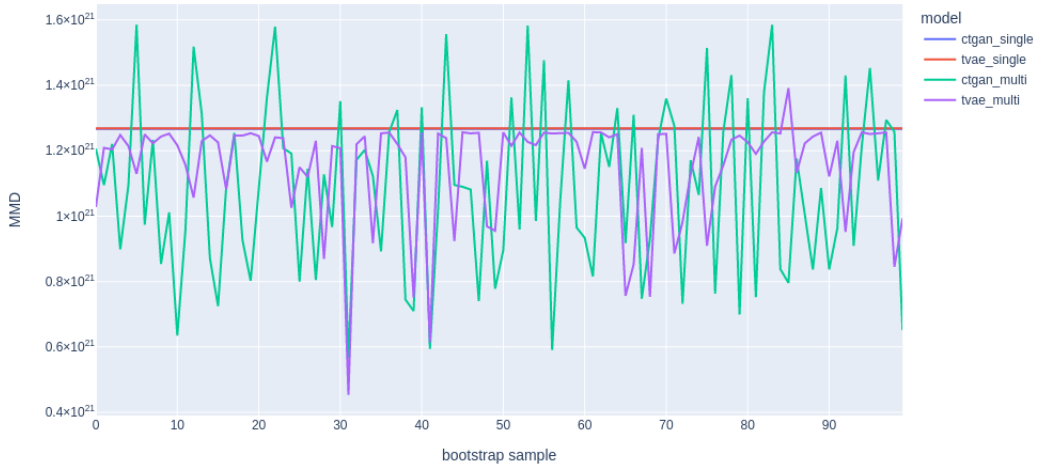
Σχήμα 5.27: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 2



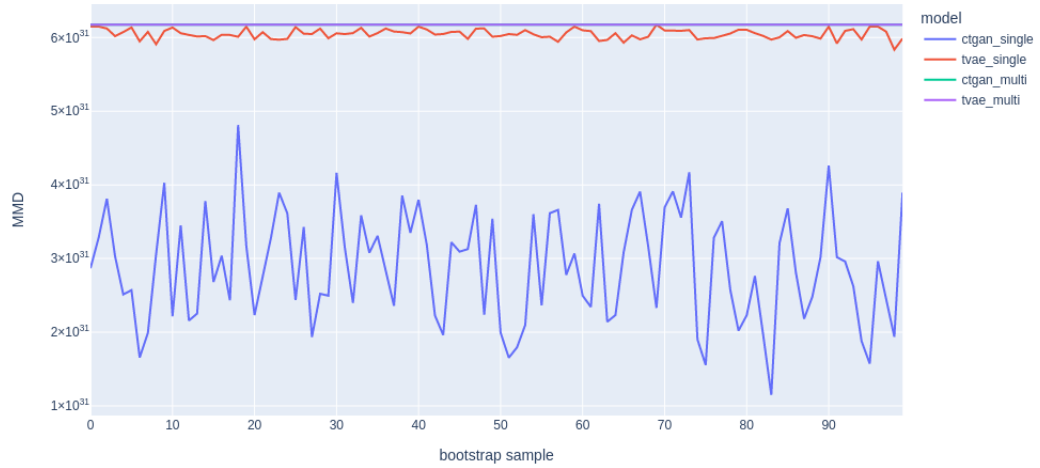
Σχήμα 5.28: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 3



Σχήμα 5.29: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 4



Σχήμα 5.30: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 5



Σχήμα 5.31: Διάγραμμα σύγκρισης των MMD των τεσσάρων μοντέλων για το προφίλ 6

5.3.2.2 Συμπεράσματα

Μερικά συμπεράσματα από τα αποτελέσματα είναι:

- Τα μοντέλα που εκπαιδεύτηκαν στα πολλά δεδομένα, όχι μόνο δεν ήταν χειρότερα από αυτά που εκπαιδεύτηκαν σε κάθε προφίλ ξεχωριστά, αλλά στις περισσότερες περιπτώσεις είχαν και καλύτερα αποτελέσματα. Αυτό αποδεικνύει ότι τα μοντέλα αυτά είναι βελτιστοποιημένα για τέτοιου τύπου εκπαίδευση (δηλαδή πάνω σε πολλά δεδομένα με διάφορες κλάσεις και όχι σε δεδομένα χωρίς κλάσεις).

- Στα μόνα δύο προφίλ που παρατηρήθηκαν λίγο χειρότερα αποτελέσματα από τα νέα μοντέλα (5.28, 5.31), πρέπει να διευκρινηστεί ότι πρόκειται για τα προφίλ με το μικρότερο ποσοστό flows (επί του συνολικού) εκ των 6 προφίλ που μελετώνται. Για αυτό τον λόγο τα μοντέλα που εκπαιδεύτηκαν στο σύνολο των δεδομένων προφανώς δυσκολεύτηκαν να μάθουν χαρακτηριστικά της κατανομής των δύο αυτών προφίλ και εξού και τα χειρότερα αποτελέσματα.

- Τα νέα μοντέλα φαίνεται να έχουν μεγαλύτερο variance από τα προηγούμενα και αυτό είναι συνέπεια του γεγονότος ότι δεν γίνεται overfitting σε ένα μόνο προφίλ και τα δεδομένα έχουν πιο διάσπαρτες τιμές.

- Μεταξύ των δύο μοντέλων (ctgan_multi, tvae_multi) δεν υπάρχει κάποιος ξεκάθαρος ΄νικητής΄ αφού το καθένα είναι καλύτερο σε διαφορετικά προφίλ (π.χ. το CTGAN στα 1, 4 και ο TVAE στα 2, 3). Αυτό δείχνει ότι για να αποφασιστεί ποιο μοντέλο θα επιλεγεί, χρειάζεται περαιτέρω μελέτη, είτε σε περισσότερα προφίλ είτε με περισσότερες παραλλαγές των παραμέτρων των μοντέλων.

Παρακάτω παρουσιάζεται και ένα αναλυτικό πινακάκι 5.3 με τους μέσους όρους και τις αποκλίσεις του MMD για τα 5 μοντέλα που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο (εκ των οποίων τα 3 πρώτα είναι per profile, ενώ τα 2 τελευταία στο σύνολο των δεδομένων), όπου αυτά είναι:

- Μοντέλο 1: GMM όπως περιγράφεται στο [19] για n_components από 2 μέχρι 30 και επιλογή του καλύτερου με BIC

- Μοντέλο 2: CTGAN με υπερπαραμέτρους όπως embedding dimension 128, generator/discriminator dimension (256,256), generator/discriminator learning rate for Adam Optimizer $2 * 10^{-4}$ μαζί με άλλες που περιγράφονται και στο [20] (με εξαίρεση το batch_size όπου δοκιμάστηκε για 2000 αντί για 500), για 150 εποχές.

- Μοντέλο 3: TVAE με υπερπαραμέτρους όπως embedding dimension 128, compress/ decompress dimensions (128,128), μαζί με άλλες που περιγράφονται και στο [21] (με εξαίρεση το batch_size όπου δοκιμάστηκε για 2000 αντί για 500), για 150 εποχές.

- Μοντέλο 4: CTGAN με υπερπαραμέτρους όπως embedding dimension 128, generator/discriminator dimension (256,256), generator/discriminator learning rate for Adam Optimizer $2 * 10^{-4}$ μαζί με άλλες που περιγράφονται και στο [20], για 200 εποχές.

- Μοντέλο 5: TVAE με υπερπαραμέτρους όπως embedding dimension 128, compress/ decompress dimensions (128,128), μαζί με άλλες που περιγράφονται και στο [21], για 200 εποχές.

Profile Number	Model Number	MMD Mean	MMD Variance
Profile 1	Model 1	$2.29 * 10^{22}$	$1.41 * 10^{45}$
	Model 2	$5.11 * 10^{22}$	$1.4 * 10^{40}$
	Model 3	$4.95 * 10^{22}$	$4.75 * 10^{40}$
	Model 4	$4.83 * 10^{22}$	$5.78 * 10^{41}$
	Model 5	$5.16 * 10^{22}$	$1.62 * 10^{40}$
Profile 2	Model 1	$5.4 * 10^{19}$	$5.71 * 10^{39}$
	Model 2	$3.64 * 10^{19}$	$2.72 * 10^{36}$
	Model 3	$5.19 * 10^{19}$	$1.79 * 10^{36}$
	Model 4	$1.22 * 10^{20}$	$3.57 * 10^{39}$
	Model 5	$5.39 * 10^{19}$	$1.21 * 10^{38}$
Profile 3	Model 1	$6.1 * 10^{21}$	$1.69 * 10^{44}$
	Model 2	$1.07 * 10^{22}$	$4.53 * 10^{36}$
	Model 3	$1.08 * 10^{22}$	$1.45 * 10^{37}$
	Model 4	$1.26 * 10^{22}$	$8.52 * 10^{41}$
	Model 5	$1.07 * 10^{22}$	$2.24 * 10^{39}$
Profile 4	Model 1	$3.36 * 10^{22}$	$9.15 * 10^{44}$
	Model 2	$3.09 * 10^{21}$	$9.25 * 10^{36}$
	Model 3	$3.12 * 10^{21}$	$9.77 * 10^{35}$
	Model 4	$3.03 * 10^{21}$	$1.04 * 10^{39}$
	Model 5	$3.11 * 10^{21}$	$6.75 * 10^{37}$
Profile 5	Model 1	$1.95 * 10^{22}$	$6.72 * 10^{45}$
	Model 2	$1.27 * 10^{21}$	$6.69 * 10^{32}$
	Model 3	$1.27 * 10^{21}$	$2.2 * 10^{32}$
	Model 4	$1.08 * 10^{21}$	$6.82 * 10^{40}$
	Model 5	$1.15 * 10^{21}$	$2.58 * 10^{40}$
Profile 6	Model 1	$1.36 * 10^{31}$	$2.78 * 10^{62}$
	Model 2	$2.88 * 10^{31}$	$5.48 * 10^{61}$
	Model 3	$6.04 * 10^{31}$	$4.26 * 10^{59}$
	Model 4	$6.17 * 10^{31}$	$2.45 * 10^{48}$
	Model 5	$6.17 * 10^{31}$	$1.59 * 10^{46}$

Πίνακας 5.3: Συνοπτική παρουσίαση αποτελεσμάτων με την μέθοδο MMD

Κεφάλαιο 6

Συμπεράσματα-Μελλοντικές Επεκτάσεις

6.1 Σύνοψη Αποτελεσμάτων

Στην συγκεκριμένη διπλωματική εργασία, αρχικά αναλύθηκαν πραγματικά δεδομένα νοσοκομειακού περιβάλλοντος για την εξαγωγή χρήσιμων συμπερασμάτων για τα επόμενα βήματα, στην συνέχεια πραγματοποιήθηκε συσταδοποίηση σε ομάδες νοσοκομειακών χρηστών για τον έλεγχο της εγκυρότητας τους και στο τέλος επιχειρήθηκε η παραγωγή συνθετικών δεδομένων από συγκεκριμένες νοσοκομειακές υπηρεσίες με χρήση παραγωγικών μοντέλων μηχανικής μάθησης και η αξιολόγηση τους (των δεδομένων και των μοντέλων) με διάφορες μεθόδους.

Στο στάδιο της ανάλυσης των δεδομένων (ή αλλιώς διερευνητικής ανάλυσης δεδομένων), εξήχθησαν πληροφορίες σχετικά με τις κατανομές αρκετών πεδίων των δεδομένων, μερικά εκ των οποίων ήταν απαραίτητα για το στάδιο της παραγωγής δεδομένων (όπως η διάρκεια ροής και τα bytes ροής). Ταυτόχρονα πραγματοποιήθηκε επιτυχής κατηγοριοποίηση χρηστών με γνώμονα την χρήση ή/και την μη-χρήση συγκεκριμένων νοσοκομειακών υπηρεσιών.

Στο στάδιο της συσταδοποίησης επιβεβαιώθηκε, και με τις δύο προσεγγίσεις που χρησιμοποιήθηκαν, σε μεγάλο βαθμό τόσο η ομοιότητα μεταξύ χρηστών που ανήκουν στον ίδιο νοσοκομειακό τομέα, όσο και η εγκυρότητα των πληροφοριών από τα DHCP Logs του νοσοκομείου, από τα οποία προέκυψαν οι προαναφερθέντες τομείς. Η προσέγγιση βέβαια με Wasserstein Distance ήταν αρκετά πιο χρονοβόρα και υπολογιστικά περίπλοκη, κάτι που την καθιστά ασύμφορη στην περίπτωση χρηστών με μεγάλο όγκο κίνησης.

Στο στάδιο της παραγωγής δεδομένων, αρχικά επετεύχθη η παραγωγή ορισμένων πεδίων δεδομένων (Duration, Interarrival Interval, Incoming Bytes, Outgoing Bytes) με τρία διαφορετικά μοντέλα (GMM, CTGAN, TVAE) εκπαιδευμένα σε **ξεχωριστά** προφίλ χρηστών, με το GMM να παράγει τα πιο ποικιλόμορφα δεδομένα (απεικονίζοντας και τις υψηλές τιμές των πραγματικών δεδομένων στα συνθετικά σε αντίθεση με τα άλλα δύο μοντέλα που ήταν biased προς τις χαμηλές τιμές, που απαρτίζουν την πλειονότητα των δεδομένων) σύμφωνα με τα διαγράμματα Quantile-Quantile ανά πεδίο δεδομένων, αν και τα άλλα δύο μοντέλα πέτυχαν ίδια και ίσως λίγο καλύτερα αποτελέσματα με την μετρική MMD συνδυασμένη με την μέθοδο Bootstrapping, στο σύνολο των δεδομένων. Στην συνέχεια, χάρη την δυνατότητα των παραγωγικών μοντέλων βαθιάς μηχανικής μάθησης (CTGAN, TVAE) για παραγωγή conditional δεδομένων, έγινε εκπαίδευση σε πολλαπλά προφίλ χρηστών **ταυτόχρονα** και τα αποτελέσματα ξεπέρασαν τις προσδοκίες, αφού με την μετρική MMD ξεπέρασαν τα αντίστοιχα μοντέλα του προηγούμενου βήματος σε απόδοση (με οριακά χειρότερα αποτελέσματα μόνο στα προφίλ με τα λιγότερα δεδομένα ροής). Αυτό απέδειξε, συν τοις άλλοις, και την χρησιμότητα των

μοντέλων αυτών σε προβλήματα με διάφορες κλάσεις (που απαιτούν conditional modeling).

6.2 Μελλοντικές Επεκτάσεις

Το σύνολο δεδομένων, το οποίο υπέστη αρκετές αλλαγές στα διάφορα βήματα της διπλωματικής εργασίας, μπορεί να χρησιμοποιηθεί και για άλλους σκοπούς, εκτός από αυτούς που μελετώνται, όπως προβλήματα ταξινόμησης (π.χ. των διαφόρων κλάσεων/χρηστών που προέκυψαν κατά την ανάλυση). Επίσης, η παρούσα διπλωματική μπορεί να χρησιμοποιηθεί ως βάση για την παραγωγή δεδομένων από περισσότερες υπηρεσίες και αυτά τα δεδομένα να απαρτίζονται από περισσότερα πεδία από αυτά που μελετήθηκαν στην διπλωματική. Αυτό μπορεί να γίνει με την περαιτέρω επεξεργασία των δεδομένων. Επιπλέον, μπορεί να γίνει επέκταση των μοντέλων που χρησιμοποιούνται, εκπαιδεύοντας τα με νέες παραμέτρους ή και με προσθήκη νέων υλοποιήσεων για την βελτίωση των παραγόμενων δεδομένων. Τέλος, η πιο προφανής επέκταση αυτής της διπλωματικής είναι αυτή που αποτέλεσε κίνητρο για την εκπόνηση της, η οποία είναι η προσομοίωση δικτυακής συμπεριφοράς τυχαίων χρηστών σε επιλεγμένες υπηρεσίες, αλλά και η αναγνώριση πραγματικών από περίεργων/επικίνδυνων συμπεριφορών/κίνησης.

Βιβλιογραφία

[1] "2020 data breach investigations report," https://www.verizon.com/about/news/verizon-2020-data-breach-investigations-report, Accessed: 2021-07-25.

[2] "2021 data breach investigations report," https://www.verizon.com/business/resources/reports/dbir/, Accessed: 2022-01-10.

[3] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[4] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

[5] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Management Science*, vol. 6, no. 4, pp. 366–422, 1960.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," 2017.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, "Improved training of wasserstein gans," 2017.

[9] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," 2014.

[10] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, 2019.

[11] Zinan Lin, Ashish Khetan, Giulia C. Fanti, and Sewoong Oh, "Pacgan: The power of two samples in generative adversarial networks," *CoRR*, vol. abs/1712.04086, 2017.

[12] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," 2014.

[13] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Mar. 2012.

[14] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley, "Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nyström method, and use of kernels in machine learning: Tutorial and survey," 2021.

[15] Rosie Dunford, Quanrong Su, and Ekraj Tamang, "The pareto principle," *The Plymouth Student Scientist*, vol. 7, pp. 140–148, 2014.

[16] Thomas M. Chen, *Network Traffic Modeling*, pp. 326–339, John Wiley & Sons, Ltd.

[17] Carl Nuzman, Iraj Saniee, Wim Sweldens, and Alan Weiss, "A compound model for tcp connection arrivals for lan and wan applications," *Computer Networks*, vol. 40, pp. 319–337, 10 2002.

[18] Anja Feldmann, "Characteristics of tcp connection arrivals," 1998, Wiley.

[19] "Gmm model api reference," https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html, Accessed: 2021-1-12.

[20] "Ctgan model api reference," https://sdv.dev/SDV/api_reference/tabular/api/sdv.tabular.ctgan.CTGAN.html#sdv.tabular.ctgan.CTGAN, Accessed: 2022-02-20.

[21] "Tvae model api reference," https://sdv.dev/SDV/api_reference/tabular/api/sdv.tabular.ctgan.TVAE.html#sdv.tabular.ctgan.TVAE/, Accessed: 2022-02-20.