|  |
| --- |
| האוניברסיטה הפתוחה  המחלקה למתמטיקה ומדעי המחשב |
| זיהוי מנגנונים ברשתות לימוד מקוונות |
| עבודת תזה זו הוגשה כחלק מהדרישות לקבלת תואר  מוסמך למדעים M.Sc. במדעי המחשב  באוניברסיטה הפתוחה  החטיבה למדעי המחשב |
|  |
| על ידי  **מאזוז משה** |
|  |

העבודה הוכנה בהנחייתו של פרופסור ראובן אביב

|  |
| --- |
|  |

אוקטובר 2009

**תוכן עניינים**

[תקציר 6](#_Toc243319153)

[1 מבוא 7](#_Toc243319154)

[2 רקע על תחום מחקר הרשתות 10](#_Toc243319155)

[2.1 מושגים כלליים בתחום הרשתות 10](#_Toc243319156)

[2.2 תכונות של רשתות 11](#_Toc243319157)

[2.2.1 עולם קטן - Small World 11](#_Toc243319158)

[2.2.2 טרנזיטיביות או Clustering 11](#_Toc243319159)

[2.2.3 התפלגות דרגות - Degree Distribution 13](#_Toc243319160)

[2.2.4 עמידות של רשתות - Network resilience 14](#_Toc243319161)

[2.2.5 עירוב תבניות – Mixing patterns 15](#_Toc243319162)

[2.2.6 תיאום דרגות – Degree correlations 16](#_Toc243319163)

[2.2.7 מבנה קהילתי – Community structure 16](#_Toc243319164)

[2.2.8 Betweens Centrality 17](#_Toc243319165)

[2.3 מודל הגרף האקראי 17](#_Toc243319166)

[2.3.1 הגדרת המודל 18](#_Toc243319167)

[2.3.2 השפעת הסתברות הקישור בין קודקודים על מבנה הגרף 18](#_Toc243319168)

[2.3.3 מודל הגרף האקראי לעומת רשתות אמפיריות 19](#_Toc243319169)

[2.3.4 הרחבות של מודל הגרף האקראי – מודל הקונפיגורציה 20](#_Toc243319170)

[2.4 רשתות "עולם קטן" 20](#_Toc243319171)

[2.4.1 תכונות המודל 21](#_Toc243319172)

[2.5 מודלים בעלי מנגנון "קישור מועדף" 21](#_Toc243319173)

[2.5.1 המודל של Price 22](#_Toc243319174)

[2.5.2 המודל של Albert & Barabasi 24](#_Toc243319175)

[2.6 מודלים להעתקה ושינוי של קודקודים 28](#_Toc243319176)

[2.6.1 מנגנון קישור לקשתות 28](#_Toc243319177)

[2.6.2 מנגנון העתקת קודקודים 29](#_Toc243319178)

[2.6.3 מנגנון הליכה על הרשת 30](#_Toc243319179)

[3 מטרת המחקר 31](#_Toc243319180)

[4 בסיס המידע 32](#_Toc243319181)

[5 מודלים ליצירת רשתות 33](#_Toc243319182)

[5.1 מודלים לבחינת קישוריות הרשתות 33](#_Toc243319186)

[5.1.1 מודל הגרף האקראי המכוון 33](#_Toc243319187)

[5.1.2 מודל "קישור מועדף" סטטי 34](#_Toc243319188)

[5.1.3 מודל "קישור מועדף" דינמי 35](#_Toc243319189)

[5.1.4 מודל "קישור מועדף" דינמי משופר 37](#_Toc243319190)

[5.1.5 מודל העתקה דינמית 38](#_Toc243319191)

[5.1.6 מודל "העולם הקטן" – גרסה מכוונת 39](#_Toc243319192)

[5.2 מודלים לבחינת עוצמת הקשרים ברשתות 40](#_Toc243319193)

[5.2.1 מודל הגרף האקראי הממושקל המכוון 40](#_Toc243319194)

[5.2.2 מודל "קישור מועדף" סטטי ממושקל 41](#_Toc243319195)

[5.2.3 מודל "קישור מועדף" דינמי ממושקל 41](#_Toc243319196)

[5.2.4 מודל "קישור מועדף" דינמי ממושקל משופר 41](#_Toc243319197)

[5.2.5 מודל "קישור מועדף" סטטי דינמי ממושקל 42](#_Toc243319198)

[6 מתודולוגיה 43](#_Toc243319199)

[6.1 בניית מסווגים 43](#_Toc243319200)

[6.2 בדיקת עמידות המסווגים 44](#_Toc243319201)

[6.3 בחירת המסווגים 44](#_Toc243319202)

[6.4 מציאת המודל הזוכה 44](#_Toc243319203)

[6.5 בדיקת איכות תוצאות הסיווג 45](#_Toc243319204)

[7 תוצאות 47](#_Toc243319205)

[7.1 תוצאות בחינת קישוריות הרשתות 47](#_Toc243319206)

[7.1.1 ביצוע הסיווג 47](#_Toc243319207)

[7.1.2 בדיקת עמידות המסווגים 51](#_Toc243319208)

[7.1.3 זיהוי המודל המתאים ביותר 53](#_Toc243319209)

[7.1.4 בדיקת איכות תוצאות הסיווג 55](#_Toc243319210)

[7.2 בחינת עוצמת הקשרים – Weighted Case 57](#_Toc243319211)

[7.2.1 ביצוע הסיווג 57](#_Toc243319212)

[7.2.2 זיהוי המודל בעל ההתאמה הטובה ביותר 60](#_Toc243319213)

[7.2.3 בדיקת עמידות המסווגים 62](#_Toc243319214)

[7.2.4 בדיקת איכות תוצאות הסיווג 64](#_Toc243319215)

[8 דיון ומסקנות 66](#_Toc243319216)

[9 סיכום וכיוונים להמשך 69](#_Toc243319217)

[נספח א' - שיטת Middendorf לייצוג רשתות מורכבות במרחב תכונות רב מימדי 71](#_Toc243319218)

[תיאור השיטה 71](#_Toc243319219)

[יתרונות השיטה 73](#_Toc243319220)

[חסרונות השיטה 73](#_Toc243319221)

[נספח ב' - תיאור חבילת התוכנה ששימשה לביצוע הבדיקות 74](#_Toc243319222)

[ב.1. מחולל המאפיינים ( Attribute Creator ) 74](#_Toc243319223)

[מבנה 74](#_Toc243319224)

[הרצת מחולל המאפיינים 82](#_Toc243319225)

[תיאור קבצי הרשתות 83](#_Toc243319226)

[תיאור התוצאות 83](#_Toc243319227)

[ב.2. מסווג הרשתות ( Networks Classifier ) 84](#_Toc243319228)

[מבנה 84](#_Toc243319229)

[הרצת מסווג הרשתות 87](#_Toc243319230)

[תיאור התוצאות 88](#_Toc243319231)

[ב.3. בודק המובהקות ( Significance Finder ) 90](#_Toc243319232)

[מבנה 90](#_Toc243319233)

[הרצת בודק המובהקות 92](#_Toc243319234)

[תיאור התוצאות 93](#_Toc243319235)

[מקורות 94](#_Toc243319236)

**רשימת טבלאות**

[טבלה 1א – אחוזי דיוק ממוצעים של המסווגים עבור המקרה הבינארי 49](#_Toc243057430)

[טבלה 2 – עמידות המסווגים עבור המקרה הבינארי 52](#_Toc243057431)

[טבלה 3א – ריכוז תוצאות סיווג הרשתות האמפיריות למודלים. 53](#_Toc243057432)

[טבלה 3ב – תוצאות הסיווג המפורטות של רשתות למודלים 54](#_Toc243057433)

[טבלה 4- נתונים מפורטים של הסתברות תוצאות הסיווג בהתאם לכמות השגיאות המירבית. 56](#_Toc243057434)

[טבלה מס' 5- כמות המסווגים שנבדקו במסגרת בדיקת המובהקות 56](#_Toc243057435)

[טבלה 6א - אחוזי דיוק ממוצעים של המסווגים עבור המקרה הממושקל 58](#_Toc243057436)

[טבלה 6ב – אחוזי דיוק של המסווגים הטובים ביותר עבור המקרה הממושקל 59](#_Toc243057437)

[טבלה 7 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים. 60](#_Toc243057438)

[טבלה 8 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים, כולל המודל המשולב. 61](#_Toc243057439)

[טבלה 9 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים, כולל שינויי ערכי הפרמטרים של מודל ה-WSPA. 61](#_Toc243057440)

[טבלה 10 – עמידות המסווגים עבור המקרה הממושקל 63](#_Toc243057441)

[טבלה 11: נתונים מפורטים של הסתברות תוצאות הסיווג בהתאם לכמות השגיאות המירבית. 65](#_Toc243057442)

[טבלה מס' 12- כמות המסווגים שנבדקו במסגרת בדיקת המובהקות 65](#_Toc243057443)

**רשימת תרשימים**

[תרשים 1 – תופעת מעבר הפאזה בגרפים אקראיים 19](#_Toc243057453)

[תרשים 2 – היסטוגרמת עמידות המסווגים עבור המקרה הבינארי 51](#_Toc243057454)

[תרשים 3- הקשר בין כמות טעויות הסיווג המירבית לבין הסתברות תוצאות הסיווג עבור המסווגים השונים. 55](#_Toc243057455)

[תרשים 4 - היסטוגרמת עמידות המסווגים עבור המקרה הממושקל 62](#_Toc243057456)

[תרשים 5- הקשר בין כמות טעויות הסיווג המירבית לבין הסתברות תוצאות הסיווג עבור המסווגים השונים במקרה הממושקל. 64](#_Toc243057457)

# תקציר

רשתות לימוד מקוונות משמשות ככלי לבניית ידע. רשתות אלו נוצרות על ידי אינטראקציה בין אנשים השייכים לקהיליית לומדים המקושרים באמצעות האינטרנט. מטרתה של עבודת תזה זו לזהות את המנגנונים אשר משפיעים על היווצרותן של רשתות אלו. במסגרת העבודה נוצר ייצוג של הרשתות באמצעות ווקטור מאפיינים רב מימדי, ונעשה שימוש בכלים מתחום "לימוד מכונה" ( Machine Learning ) על מנת לגלות מי מבין מגוון מודלים סטוכסטיים ליצירת רשתות הינו המתאים ביותר ליצירת רשתות הלמידה המקוונות. בעבודה זו נבדקו שני היבטים של הרשתות: הקישוריות בין הקודקודים ועוצמת הקשרים בין הקודקודים. תוצאות הבדיקה מראות כי במקרה של בחינת הקישוריות ברשת, המנגנון המשפיע על יצירת הרשתות המקוונות הינו מסוג "צימוד מועדף" סטטי ( Preferential Attachment ), כלומר, המשתתפים ברשת מעדיפים ליצור קשרים עם משתתפים שהינם בעלי משקל גבוה, כאשר משקל זה נקבע בזמן יצירת הרשת. בחינת עוצמת הקשרים ברשת מגלה כי גם במקרה זה המנגנון המשפיע על יצירת הרשתות הינו מסוג "צימוד מועדף" סטטי, אם כי במקרה זה ההתאמה של המודל מעט חלשה יותר, וייתכן כי מנגנונים נוספים מתחילים להיכנס לפעולה ולהשפיע על מבנה הרשת.

חלקים מתוצאות עבודת תזה זו הוצגו בכנסים הבאים:

1. **Chais 2009** – האדם הלומד בעידן הטכנולוגי, כנס צ'ייס השנתי הרביעי למחקרי טכנולוגיות למידה, האוניברסיטה הפתוחה, פברואר 2009.
2. **WEB 2009**- The Eighth IASTED International Conference on Web-based Education, פוקט תאילנד, מרץ 2009.

# מבוא

תופעות רבות במגוון תחומים ניתנות לייצוג באמצעות רשתות, כגון: רשתות חברתיות, רשתות ביולוגיות, רשת האינטרנט, רשת ה-WWW, רשתות ציטוטי מאמרים ועוד. בעשור האחרון התרחשו מספר התפתחויות אשר השפיעו על תחום מחקר הרשתות:

* חל גידול עצום הן בתחומים בהם נאסף מידע הניתן לייצוג כרשתות והן בגודל ובמורכבות של הרשתות. בין תחומים אלו ניתן למנות רשתות המבוססות על התפתחות רשת האינטרנט, וכן רשתות המבוססות על ההתפתחויות בתחום הביולוגיה אשר כללו בין השאר את מיפוי הגנום של יצורים שונים ובכללם הגנום האנושי. כתוצאה מכך כמות הרשתות הזמינות למחקר גדלה באופן משמעותי.
* ההתפתחות המתמדת בכוח המחשוב אפשרה יצירה ושימוש בכלים בעלי עוצמה רבה המאפשרים לחקור רשתות אלו. ניתוח אנליטי של הרשתות איננו תמיד אפשרי עקב גודלן ומורכבותן של הרשתות או כתוצאה מהמורכבות של המודלים המתיימרים לייצג את הרשתות. במקרים אלו ניתן לבצע שימוש בסימולציות ממוחשבות של הרשתות וכך לבדוק את התכונות שלהן מחד, ולהגיע לתובנות על המתרחש בהן מאידך.

התפתחויות אלו וההבנה כי תכונות הרשת מושפעות בעיקר מאופי הקישוריות הובילו לניסיונות לנתח את תכונות הרשתות. השוואת תכונות הרשתות האמפיריות לרשתות הנוצרות באמצעות שימוש במודל הגרף האקראי הקלאסי (Erdos & R'enyi, 1960 ), שהיה מקובל בעולם מחקר הרשתות, העלתה כי במקרים רבים התכונות של הרשתות האמפיריות שונות מהתכונות המאפיינות רשתות אשר מחוללות באמצעות מודל זה.

בין תכונות אלו ניתן למנות את ( Newman, 2003 ) :

* **התפלגות דרגות הקודקודים בגרף.**
* **ערך מקדם ה-Clustering**: הסיכוי ששני צמתים המקושרים לצומת שלישי מקושרים ביניהם.

כתוצאה מכך החל מחקר ענף בתחום הרשתות אשר התמקד במספר היבטים וביניהם:

1. מציאת הערכים של תכונות מוכרות וגילוי תכונות חדשות ברשתות האמפיריות.
2. זיהוי תתי-מבנים לא טריוויאליים ( Motifים ) בתוך הרשתות האמפיריות.
3. זיהוי מנגנונים המשפיעים על היווצרות הרשתות.
4. בדיקת השפעת מבנה הרשת על תהליכים המתרחשים בה.

עבודת תזה זו עוסקת בתחום זיהוי המנגנונים המשפיעים על היווצרות הרשתות. תחום זה מבוסס על פיתוח מודלים ליצירת רשתות אשר מנסים לחקות את התכונות של הרשתות האמפיריות. בשנים האחרונות הוצע מגוון נרחב של מודלים לבניית רשתות. מודלים אלו מבוססים על מנגנונים שונים ומכילים מרכיב סטוכסטי. המודלים הוצעו ע"י חוקרים שונים כגון: Watts & Strogatz ( 1998), Krapivsky et al ( 2002), Vazquez ( 2002), Shen-Orr ( 2002 ) ורבים נוספים.

על מנת לבדוק האם מודל ליצירת רשתות אכן מצליח לחקות את תכונות הרשתות האמפיריות, מבוצעת בדיקת התאמה בין המודל המוצע לרשתות האמפיריות. השיטה המקובלת לביצוע בדיקה זו מבוססת על השוואת מספר מאפיינים מרכזיים של הרשת למאפיינים של רשתות הנבנות באמצעות המודלים. אולם גישה זו סובלת משני חסרונות מהותיים:

1. מציאת המודל נעשית ע"פ בדיקת קבוצת מאפיינים מצומצמת אשר נבחרה מראש ולכן עשויה להיות מוטה למטרות הבדיקה.
2. עקב ריבוי המודלים השונים, ייתכן כי תוצאות בדיקת ההתאמה ייראו כי מספר מודלים מספקים את התכונות הנדרשות, ברמת דמיון שלא תאפשר אבחנה ביניהם.

עקב כך, הסתמכות על מודל שהראה התאמה בשיטה זו למטרת ניתוח הרשת הינו בעייתי הואיל ומודלים שונים המבוססים על מנגנונים השונים מהותית זה מזה, עשויים לתת התאמה דומה.

על מנת להתגבר על בעיה זו הוצעה ע"י Middendorf et al ( 2004 ) שיטה כללית לסיווג רשתות, המבוססת על טכניקות של כריית מידע. שיטה זו מבוססת על שני מרכיבים עיקריים:

1. יצירת ווקטור מאפיינים רב מימדי עבור כל רשת באופן שיטתי.
2. שימוש במסווגים ( Classifiers ) לשיוך הרשת האמפירית לאחד המודלים.

רשתות לימוד מקוונות נוצרות על ידי אינטראקציה בין אנשים המקושרים באמצעות האינטרנט. הרשתות נוצרות כתוצאה משיתוף פעולה בקהיליית הלומדים באמצעות מתן מענה להודעות המפורסמות ע"י חברי הקהילה. עבודות קודמות אשר בדקו רשתות אלו (Aviv et al., 2007a, Aviv et al., 2007b ) זיהו את קיומם של מבנים לא טריוויאליים ברשתות ( Motifים ). כמו-כן, עבודות אלו זיהו כי רשתות אלו אינן מתאימות למודל הגרף האקראי, אולם ייתכן שהן מתאימות למודלים אחרים מורכבים יותר.

**בתזה זו נעשה שימוש בשיטת הסיווג שהוצעה ע"י Middendorf (2004) לניתוח רשתות לימוד מקוונות על מנת למצוא את המודל המתאים ביותר להתפתחות רשתות אלו, ובכך לנסות ולהתחקות אחר המנגנון העומד בבסיס התופעה.**

המבנה של עבודת תזה זו הוא כדלהלן: רקע כללי על נושאי המחקר בתחום הרשתות האקראיות מתואר בחלק 2. ניסוח מטרת המחקר מתוארת בחלק 3. בחלקים 4 ו-5 מתואר בסיס המידע והמודלים ליצירת רשתות בהם בוצע שימוש במסגרת המחקר, בהתאמה. בחלק 6 מתוארת המתודולוגיה אשר שימשה לבדיקת וניתוח התוצאות, ובחלק 7 מתוארות התוצאות. דיון ומסקנות מופיעים בחלק 8, ובחלק 9 מופיע סיכום וכיוונים להמשך מחקר עתידי .

# רקע על תחום מחקר הרשתות

תחום מחקר הרשתות הינו תחום מאוד נרחב אשר זכה לפריחה בשנים האחרונות. בחלק זה מובאת סקירה קצרה של נושאי המחקר בתחום הרשתות. הסקירה עוסקת בתכונות של רשתות אשר נמצאות במוקד המחקר וכן במודלים המרכזיים ליצירת רשתות אקראיות שפותחו במטרה לדמות את הרשתות האמפיריות. הסקירה מבוססת, בעיקרה, על מספר סקירות מקיפות אשר פורסמו בשנים האחרונות: Newman ( 2003 ), Barabasi ( 2002 ), Buts ( 2008 (.

## מושגים כלליים בתחום הרשתות

לאורך העבודה ייעשה שימוש במספר מושגים מתחום הרשתות. להלן מובא תיאור קצר של המושגים השונים והסימונים בהם ייעשה שימוש. במידת הצורך יובא תיאור נרחב יותר של המושג במקומות הרלבנטיים.

1. **רשת - Network:** קבוצה של עצמים המכונים קודקודים (Nodes,Vertices) עם קשרים ביניהם המכונים קשתות ( Edges ). **(** נקראת גם **"גרף" ( Graph )** בספרות המתמטית ).
2. **N:** מספר הקודקודים ברשת.
3. **מסלול (Path) בין קודקוד לקודקוד :** קבוצת סדורה של קשתות, אשר ניתן להגיע באמצעותן מקודקוד לקודקוד .
4. **אורך מסלול--Length:** גודל קבוצת הקשתות המרכיבות את המסלול. מסומן ב-ℓ**. מרחק בין קודקודים-Distance:** האורך המינימאלי של מסלול בין הקודקודים. מסומן ב-d.
5. **קליקה – Clique:** קבוצת קודקודים אשר כל זוג מהם מחובר ישירות ע"י קשת.
6. **רשת לא מכוונת-Undirected Network/Graph:** רשת בה לא מוגדר כיוון לקשתות.
7. **רשת מכוונת-Directed Network/Graph-:** רשת בה עבור כל קשת מוגדר כיוון, כך שיש לקשת קודקוד ממנו היא יוצאת וקודקוד אליו היא נכנסת.
8. **רשת ממושקלת – Weighted Network/Graph:** רשת בה מוגדר משקל לכל אחת מהקשתות.
9. **רכיב קשירות -Component:** קבוצת קודקודים מקסימאלית, אשר בין כל שניים מהם קיים מסלול.
10. **רכיב קשירות חזק** - **Strongly connected component**: רכיב קשירות ברשת מכוונת בו ישנו מסלול מכוון בין כל זוג קודקודים.
11. **רשת לא קשירה:** רשת בה ישנם מספר רכיבי קשירות.
12. **דרגת הקודקוד- Degree:** כמות הקודקודים ברשת המקושרים ישירות לקודקוד.
13. **דרגה יציאה של קודקוד-OutDegree:** ברשת מכוונת, כמות קשתות היוצאות מהקודקוד.
14. **דרגה כניסה של קודקוד- InDegree:** ברשת מכוונת, כמות הקשתות הנכנסות לקודקוד.
15. **מסלול גיאודזי-Geodesic Path:** המסלול הקצר ביותר בין זוג קודקודים ברשת.
16. **רכיב ענק-Giant Strong Component:** רכיב קשירות חזק ברשת אשר מקיים את התנאים הבאים: מספר הקודקודים שהוא מכיל פרופורציונאלי לכמות הקודקודים ברשת, ולא קיימים רכיבי קשירות ברשת המכילים את אותו סדר גודל של קודקודים מלבדו.
17. **רשת חסרת מידה** – Scale Free Network: רשת בעלת התפלגות דרגות קודקודים מעריכית. ( ראה פירוט בסעיף 2.2.3 ).

## תכונות של רשתות

תכונות רבות של רשתות נחקרו במהלך השנים האחרונות. בין התכונות שזכו להיות במוקד המחקרים ניתן למנות את התכונות הבאות:

### עולם קטן - Small World

בשנות ה-60 של המאה ה-20 נערך ניסוי מפורסם ע"י מילגרם (Milgram, 1967). במהלך ניסוי זה התבקשו סטודנטים להעביר מכתבים למען מסוים, אשר לא היה מוכר להם. המכתבים הועברו באמצעות העברתם לאנשים אחרים, המוכרים להם אישית, אשר לדעתם יוכלו לבצע את המשימה טוב יותר. רוב המכתבים אבדו במהלך הניסוי אולם כרבע הגיעו ליעדם ועברו בממוצע דרך כששה אנשים בלבד. תופעה זו זכתה לכינוי "עולם קטן" ( Small World ) ובאופן כללי היא התכונה המאפיינת רשתות בהן המרחק הממוצע בין כל זוגות הקודקודים הינו קטן יחסית לכמות הקודקודים ברשת. עקב חוסר היכולת למדוד מרחק ממוצע בין קודקודים ברשתות לא קשירות, נהוג להתעלם בחישוב זה מזוגות קודקודים השייכים לרכיבי קשירות שונים. באופן פורמאלי, מייחסים את תכונת ה-Small-World לרשתות בהן המרחק הממוצע בין כל זוגות הקודקודים גדל באופן לוגריתמי יחסית לכמות הקודקודים ברשת. עבור רשתות "חסרות מידה" ( Scale Free ), הוכח (Bollob´as & Riordan, 2002 ) חסם עליון על יחס זה שערכו , כאשר N הינו מספר הקודקודים ברשת.

### טרנזיטיביות או Clustering

ברשתות רבות התגלה כי אם קודקוד A מקושר ישירות לקודקוד B, וקודקוד B מקושר ישירות לקודקוד C אזי קיימת הסתברות מוגברת שקיים קישור ישיר בין קודקוד A לקודקוד C. מבחינה טופולוגית ניתן לומר שתכונת הטרנזיטיביות באה לידי ביטוי בקיום "משולשים" ברשת – קליקות בגודל 3. במילים אחרות, ערך מקדם ה-Clustering הינו ההסתברות ששני קודקודים שהינם שכנים של קודקוד שלישי הם גם שכנים בעצמם. ערך מקדם ה-Clustering ניתן להגדרה הן ברמת הקודקוד והן ברמת הרשת.

* **ברמת הקודקוד הבודד**, ערך מקדם ה-Clustering מוגדר כמספר הקשרים בין שכני הקודקוד חלקי כלל הקשרים האפשריים בין השכנים, כאשר עבור קודקודים שדרגתם 0 או 1, מוגדר ערך מקדם ה-clustering שלהם להיות 0.
* **ערך מקדם ה-Clustering של רשת** ניתן להגדרה בשלוש דרכים שאינן שקולות:
  1. ממוצע ערך מקדם ה-Clustering של קודקודי הרשת.
  2. אחוז ה"משולשים" ברשת. או ביתר דיוק, 3 כפול מספר המשולשים ברשת חלקי כמות השלשות של קודקודים בהן לפחות קודקוד אחד קשור לשני הקודקודים האחרים. הסיבה להכפלה ב-3 נובעת מכך שכל שלשה של קודקודים בהן כל קודקוד מקושר למשנהו נספרת כשלוש שלשות של קודקודים בהם לפחות קודקוד אחד קשור לשני הקודקודים האחרים. הכפלה זו מאפשרת לנרמל את ערך מקדם ה-Clustering, כך שיהיה בין 0 ל-1.
  3. מספר המשולשים ברשת כפול 6 חלקי כמות המסלולים באורך 2. כמו בסעיף ב' לעיל החלוקה ב-6 נועדה לנרמל את הערך בין 0 ל-1.

ההגדרות בסעיפים ב' ו-ג' הינן שקולות, אולם ההגדרה בסעיף א' שונה מהן בכך ששימוש בה נותן תוצאות גבוהות יותר עבור קודקודים בעלי דרגה נמוכה. לכן חשוב לציין את אופן החישוב של ערך מקדם ה-Clustering על מנת למנוע טעויות.

**ערכי מקדם ה-Clustering ברשתות אמפיריות**

באופן כללי, ללא קשר לשיטת החישוב, ערך מקדם ה-Clustering של רשתות אמפיריות גדול משמעותית מערכי מקדמי ה-Clustering המתקבלים עבור רשתות המחוללות באמצעות מודל הגרף האקראי (Newman, 2002). בעוד שעבור רשתות אקראיות ערך מקדם ה-Clustering שואף ל-, הרי שברשתות אמפיריות ( או מודלים המנסים לדמות רשתות אמפיריות ) ערך מקדם ה-Clustering שואף לקבוע , ולכן הפער בין הערכים האמפיריים, לבין הערכים של מודל הגרף האקראי הוא מסדר גודל של .

**הכללת מושג ה-Clustering**

מושג ה-Clustering המקורי מתייחס לשלשות של קודקודים. אולם ניתן להכליל מושג זה ולהגדירו על מסלולים ארוכים או קצרים יותר. ההכללה הנפוצה ביותר היא עבור מסלולים באורך 2, כאשר בין זוג קודקודים עשויים להיות מוגדרות מספר קשתות. הכללה זו במקרה של גרפים מכוונים נקראת בשם **הדדיות** ( Reciprocity ) והיא נבדקת בד"כ במחקרים על רשתות חברתיות Scott, 2000)).

### התפלגות דרגות - Degree Distribution

חלוקת הקודקודים ברשת לפי דרגותיהם ומציאת הפונקציה המתאימה לתיאור חלוקה זו נקראת התפלגות דרגות הקודקודים ברשת. אם נגדיר את להיות אחוז הקודקודים ברשת שדרגתם , אזי הגדרה זו שקולה לכך כי הינה ההסתברות שבבחירת קודקוד אקראי ברשת נקבל קודקוד שדרגתו . תיאור של ערכי ה- ניתן לביצוע ע"י יצירת היסטוגרמה של דרגות הקודקודים ברשת. היסטוגרמה זו נקראת התפלגות דרגות הקודקודים ברשת.

#### מציאת פונקצית ההתפלגות

במודל הגרף האקראי כל קשת קיימת או לא קיימת בהסתברות זהה, ולכן התפלגות דרגות הקודקודים היא בינומית ( או פואסונית עבור ערכים גדולים ). אולם ברוב המקרים, רשתות אמפיריות אינן חולקות התפלגות זאת. במקרים רבים התפלגות דרגות הקודקודים ברשתות אלו מוטה באופן ניכר, כלומר לרוב הקודקודים ישנה דרגה השונה בהרבה מהממוצע. כמו-כן להתפלגות האמפירית יש "זנב" ארוך, כלומר הערכים לא מרוכזים במספר נקודות בודדות, אלא מתפרשים על פני טווח נרחב של ערכים. מדידת התפלגות הדרגות ברשתות **אמפיריות גדולות**, כגון רשת ה- WWW ( רשת הקישורים בין העמודים באינטרנט )**,** הינה בעייתית ונובעת משתי סיבות: מחד עקב גודלה של הרשת לא ניתן בד"כ לחשב את ההתפלגות המדויקת הואיל ודבר זה מחייב מעבר על כלל הרשת –דבר שאיננו מעשי ( בין השאר עקב העובדה שדרגת הקודקודים צריכה להיות מחושבת, ולעיתים מדובר על תהליך מורכב ), ומאידך עקב טווח הערכים הנרחב שימוש במדגם בד"כ לא מכיל מספיק דוגמאות על מנת להיות תקף מבחינה סטטיסטית.

קיימות שתי שיטות מקובלות לעקיפת בעיה זו:

* 1. בניית היסטוגרמה בה ישנה חלוקה ל"תרמילים" ( bins) הגדלים באופן מעריכי. על מנת לנרמל את התוצאות, כמות הקודקודים בכל תרמיל נמדדת ומחולקת ברוחב התרמיל ( כמות הערכים שה"תרמיל" מכיל ). השימוש העיקרי של שיטה זו הינו במקרה בו מציגים את התוצאות בסולם לוגריתמי, הואיל ואז גודל כל תרמיל נראה זהה. בנוסף, ככל שמגיעים ל"זנב" ההתפלגות התרמיל מתרחב ולכן הבעיות הסטטיסטיות מצטמצמות.
  2. שימוש בפונקצית התפלגות מצטברת ( cdf ). כלומר במקום להשתמש ב- משתמשים ב- . לשיטה זו מספר יתרונות: ראשית אנו לא מאבדים מידע, ושנית כמות הרעש ב"זנב" ההתפלגות פוחתת. אולם לשיטה זו ישנם גם חסרונות: לא ניתן לקבל תצוגה ישירה של ההתפלגות עצמה באמצעות ציור הגרף של הפונקציה. כמו-כן הנקודות הסמוכות בתרשים תלויות מבחינה סטטיסטית, ולכן מציאת פונקצית התאמה נכונה למידע הינה מורכבת יותר.

#### התפלגות דרגות בגרפים מכוונים

עבור גרפים מכוונים לכל קודקוד ישנה דרגת כניסה ודרגת יציאה ולכן ההתפלגות תלויה בשני ערכים. עם זאת במחקרים על רשתות מכוונות – למשל רשת ה-WWW, קיימת בד"כ התייחסות רק לדרגת היציאה או הכניסה בלבד או שישנה התעלמות מהכיוון של הקשתות (Albert & Barabasi, 1999, Broder et al, 2000). עם זאת ישנם מחקרים ( Newman et al, 2002 ) המראים כי ישנו מתאם חזק בין דרגות הכניסה לדרגות היציאה עבור רשתות מסוימות.

#### רשתות חסרות מידה - Scale Free Networks

רשתות עם התפלגות דרגות מעריכית זכו להתעניינות רבה (Strogatz, 2000 , Albert & Barabasi 2002). עבור רשתות אלו התפלגות דרגות הקודקודים מקיימת: , כאשר הינו בעל ערך חיובי. כלומר, ההסתברות שקודקוד ברשת יהיה בעל דרגה היא פרופורציונית ל- . מכך נובע שכמות הקודקודים בעלי דרגות גבוהות הולכת וקטנה באופן מעריכי. רשתות אלו נקראות רשתות "חסרות מידה" - Scale Free Networks. אולם "חוסר המידה" של רשתות אלו, כלומר חוסר התלות שלהן בכמות הקודקודים ברשת ( N) מתייחס בד"כ רק להתפלגות דרגות הקודקודים שלהן ולא לתכונות אחרות של הרשת. רשתות מסוג זה התגלו במגוון רב של מקרים ( Newman, 2002 ), כגון: רשתות ציטוטים של מאמרים, רשת האינטרנט וה-WWW, רשתות מטבוליות ועוד.

### עמידות של רשתות - Network resilience

העמידות של רשתות בפני הסרה של קודקודים קשורה להתפלגות דרגות הקודקודים ברשת. רוב הרשתות תלויות לצורך תפקודן בקישוריות שלהן - כלומר בקיום מסלולים בין הקודקודים השונים ברשת. כאשר מסירים קודקודים מהרשת, אורך המסלול בין הקודקודים הולך וגדל, עד שבשלב מסוים הרשת מתפרקת למספר רכיבי קשירות. לרשתות שונות ישנה עמידות שונה כלפי מצב זה של הסרת קודקודים אשר תלויה לא רק בכמות הקודקודים המוסרת אלא גם בקודקודים שנבחרו להסרה. ניתן למשל להסיר קודקודים באופן אקראי או להסיר קודקודים מסוג מסוים, כגון הקודקודים עם הדרגות הגבוהות ביותר. עמידות רשתות היא בעלת חשיבות למשל בהתפשטות נזקים, כגון מחלות באוכלוסיה, וירוסי-מחשב וכדומה. במקרה זה “הסרה” של קודקודים ברשת שקולה למשל לחיסון של אנשים כנגד המחלה. הואיל וחיסון של אנשים ספציפיים הינו בעל השפעה לא רק על אותם אנשים אלא עשוי להשפיע על היכולת של המחלה להתפשט לאנשים אחרים, נובע מכך כי ההחלטה את מי לחסן יכולה למנוע התפשטות מחלות ביעילות גבוהה מבלי לחסן את כלל האוכלוסייה.

#### השפעות על אורך המסלול הממוצע

מחקרים נוספים, לבדיקת עמידות רשתות, בדקו את השתנות אורך המסלול הממוצע בין קודקודים כתוצאה מהסרת קודקודים ברשת. אחד המחקרים ( Albert et al, 2000 ) הראה כי רשתות מסוג חסרות מידה ( כמו רשת האינטרנט ) הן בעלות עמידות גבוהה בפני הסרה אקראית של קודקודים,כלומר אורך המסלול הממוצע לא הושפע במידה ניכרת עקב הסרת הקודקודים, זאת מכיוון שבהסרה אקראית יוסרו, ברוב המקרים, קודקודים נפוצים, ואלו בעלי דרגה נמוכה מאוד. אולם הסרת אחוז קטן יחסית של הקודקודים בעלי הדרגות הגבוהות מובילה במהרה לשיתוק התקשורת ברשת, כלומר להתפרקות הרשת למספר רכיבי קשירות. באופן מעניין מחקר דומה (Broder et al, 2000) הראה כי על מנת לשתק את רשת האינטרנט, יש להסיר את כל הקודקודים שדרגתם חמש ומעלה – צעד אשר נראה הרבה יותר דרסטי בהתחשב בעובדה שישנם קודקודים שדרגתם גבוהה מאוד. לכן הסיקו החוקרים מסקנה הפוכה, כלומר שהרשת מאוד עמידה בפני הסרה של קודקודים בעלי דרגה גבוהה. אולם בחינה מעמיקה של התוצאות הראתה כי למעשה אין בכך סתירה. הסיבה לכך נובעת מהעובדה שברשתות חסרות מידה, כמו רשת האינטרנט, כמות הקודקודים בעלי הדרגות הגבוהות קטנה באופן מעריכי. לכן מספיק להסיר אחוז קטן של הקודקודים על מנת להסיר את כל הקודקודים שדרגתם גדולה מחמש. בעקבות מחקרים אלו בוצעו מחקרים נוספים (Jeong et al 2001, Dunne et al 2002) על רשתות חסרות מידה נוספות וגם עבורן התקבלו תוצאות דומות.

### עירוב תבניות – Mixing patterns

היבט נוסף בתחום מחקר הרשתות הוגדר ע"י Newman ( 2003 ) בשם "עירוב תבניות". "עירוב תבניות" עוסק בחקירת אופן הקישור של הקודקודים השונים ברשת. ברשתות שונות ניתן לגלות שישנם קבוצות של קודקודים, כך שהסיכוי לקיום קשרים בתוך הקבוצה הינו נמוך ואילו קשרים בין קבוצות יותר נפוצים. למשל, ברשת האינטרנט ניתן להגדיר שלושה סוגי קודקודים – משתמשי קצה, ספקי גישה וספקי תשתית, כאשר הקישורים הינם בדרך כלל בין משתמשי קצה לספקי גישה ובין ספקי גישה לספקי תשתית אך לא בתוך הקבוצות. המצב יכול להיות גם הפוך ובו בתוך הקבוצה ישנם קשרים, אולם בין הקבוצות כמעט ואין קשרים – למשל ברשתות חברתיות מרוכזים הקשרים לפי קבוצות עניין, מוצא וכדו'. תכונה זו מכונה "עירוב מסודר" ( Assortative mixing ).

### תיאום דרגות – Degree correlations

Newman ( 2003 ( הגדיר בשם "תיאום דרגות" מקרה פרטי של "עירוב תבניות". במקרה זה נחקר הקשר בין הקשרים בין קודקודים לפי הדרגות שלהם. כלומר, האם קודקודים בעלי דרגה גבוהה מקושרים גם הם לקודקודים בעלי דרגות גבוהות, או שדווקא הם מקושרים לקודקודים בעלי דרגות נמוכות. התשובה לכך איננה חד משמעית והתגלו רשתות משני הסוגים. הסיבה ש"תיאום הדרגות" זוכה למעמד בפני עצמו ולא נכלל במחקרים בנושא "עירוב תבניות", נובעת מכך שדרגת הקודקודים הינה חלק מהטופולוגיה של הרשת, ולכן יכולה להשפיע על מבנה הרשת. מספר דרכים הוצעו לבדיקת תיאום הדרגות, החל משימוש בהיסטוגרמה פשוטה (Maslov et al, 2002 ), חישוב פונקציה המתאימה בין דרגת הקודקוד לבין הדרגה הממוצעת של שכניו (Pastor-Satorras et al, 2001), וכלה בחישוב מקדם ההתאמה של Pearson בכל צד של הקשתות (Newman et al , 2002). באופן כללי התגלה כי עבור רשתות חברתיות יש תיאום דרגות, אולם עבור רשתות מסוגים אחרים ( רשתות טכנולוגיות, רשתות מידע, רשתות ביולוגיות ) אין תיאום דרגות,. הסיבה להבדל זה עדיין לא ידועה כמו-כן לא ידוע אם יש לה הסבר יחיד.

### מבנה קהילתי – Community structure

רשתות חברתיות הינן רשתות המורכבות מקבוצות של אנשים ביניהם קיימים קשרים מסוג אחד או יותר, כגון קשרים אישיים, מקצועיים ועוד. מקובל להניח כי רשתות חברתיות הינן בעלות מבנה קהילתי (Scott, 2000), (Wasserman & Faust, 1974) . רשתות בעלות מבנה קהילתי הינן רשתות המכילות קבוצות של קודקודים אשר ביניהם יש צפיפות גבוהה של קשרים, בעוד שבין הקבוצות צפיפות הקשרים נמוכה. אולם רשתות חברתיות אינן היחידות בעלות מבנה זה. מבנה קהילתי יכול להימצא גם ברשתות מסוגים אחרים. למשל ברשתות של ציטוטי מאמרים קבוצות הקודקודים יכולות להיות תחומי עניין שונים. ברשתות נוירונים הקבוצות יכולות לייצג יחידות תפקודיות וכדומה.

#### זיהוי המבנה הקהילתי

אחת הדרכים המקובלות לזיהוי המבנה הקהילתי של הרשת, הינה ע"י שימוש בטכניקות של ניתוח אשכולות (Everitt, 1974) הקרויות לעיתים בשם אשכול היררכי (Hierarchical Clustering ). בשיטה זו נותנים לכל זוג קודקודים משקל המייצג את חוזק הקשר ביניהם. באופן כללי, המשקל ניתן לכל זוג קודקודים ולא רק לקודקודים הקשורים ביניהם ע"י קשת. שיטה אחת לביצוע האשכול הינה להתחיל כאשר כל קודקוד הינו אשכול בפני עצמו ובכל שלב למזג את שני האשכולות המחוברים ע"י הקשת בעלת המשקל הגבוה ביותר. בכל שלב ניתן לבחון את התוצאה ורכיבי הקשירות המתקבלים הינם הקהילות. בסיום התהליך מתקבל רכיב קשירות אחד – "קהילה" אחת. התהליך כולו ניתן לייצוג באמצעות תרשים עץ ( Dendogram ), כאשר את הקהילות הקיימות בכל שלב ניתן לראות באמצעות ביצוע חתך אופקי על תרשים העץ. חוזק הקשר בין הקודקודים ניתן להגדרה בדרכים רבות, וביניהם: המרחקים בין הקודקודים, גודל החתך המינימאלי בין הקודקודים, משקל המסלול בין הקודקודים (בגרפים ממושקלים ) ועוד.

### Betweens Centrality

ה- Betweens Centrality של קודקוד ברשת מוגדר להיות כמות המסלולים הגיאודזיים בין קודקודים ברשת העוברים דרך קודקוד זה.Goh et al ( 2002 ) הראה כי עבור רשתות רבות, ערך ה-Betweens הינו בעל התפלגות מעריכית. עקב זאת הציעו לסווג את הרשתות לפי ערכו של המעריך בהתפלגות זו. ערך ה-Betweens יכול לשמש גם כמדד לעמידות של הרשת. במובן שהסרה של קודקוד בעל ערך Betweens centrality גבוה יאריך את המסלול בין זוגות קודקודים רבים יותר.

## מודל הגרף האקראי

הניסיון הרציני הראשון ליצור מודל אקראי עבור רשתות גדולות היה ה"רשת האקראית" של Rapoport ( 1957 ). מודל זה התגלה שוב, באופן עצמאי כעשור מאוחר יותר ע"י Erdos & R'enyi (1960), אשר חקרו אותו באופן מעמיק ומפורט, ונתנו לו את השם "גרף אקראי" ) Random Graph ( בו הוא מוכר כיום.

תכונות רבות של הרשתות הנוצרות באמצעות מודל הגרף האקראי, אינן מתאימות לתכונות חשובות שהתגלו ברשתות אמפיריות. למשל התפלגות דרגות הקודקודים במודל הגרף האקראי שונה מהתפלגות הדרגות המעריכית הקיימת ברשתות אמפיריות רבות. לכן הורחב המודל המקורי במגוון דרכים. הרחבות של המודל, המאפשרות ליצור רשתות בעלות התפלגות דרגות קודקודים שונה, הובילה ליצירת מחלקה של מודלים הידועים בשם "גרפים אקראיים מוכללים" ( Generalized Random Graphs ) או "מודל הקונפיגורציה" ( Configuration Model ).

### הגדרת המודל

מודל הגרף האקראי מוגדר באופן הבא: בוחרים את מספר הקודקודים שברצוננו שיהיו ברשת. עבור כל זוג קודקודים, מחליטים בהסתברות האם לקשר אותם באמצעות קשת. תהליך זה מגדיר גרף אקראי אשר Erdos & R'enyi סימנו באמצעות . הינו למעשה אוסף של כל הגרפים בעלי קודקודים בהם גרף בעל קשתות מופיע בהסתברות , כאשר הינו מספר הקשתות המקסימאלי בגרף. Erdos & R'enyi הגדירו מודל דומה נוסף בשם , שהינו אוסף כל הגרפים בעלי קודקודים בהם יש בדיוק קשתות. כאשר לכל אחד מהגרפים באוסף יש הסתברות שווה להופיע. תכונות רבות של הגרף האקראי ניתנות לפתרון מדויק עבור גרפים בעלי מספר קודקודים גבוה.

הואיל וקיום הקשתות במודל בלתי תלוי זו בזו, אנו מקבלים כי לכל קודקוד יש בממוצע שכנים. כמו-כן הסיכוי שקודקוד מסוים הינו בעל דרגה שווה ל- . לכן התפלגות דרגות הקודקודים היא בינומית ( או פואסונית עבור ערכים גדולים של ). כלומר , כאשר הקירוב הופך להיות מדויק עבור הגבול של גדול ו- קבוע .

### השפעת הסתברות הקישור בין קודקודים על מבנה הגרף

המבנה של הגרף האקראי משתנה בהתאם לערכו של . ככל ש- גדול יותר כך כמות הקשתות גדלה ונוצרים רכיבי קשירות גדולים יותר. אחת התכונות החשובות ביותר של גרף אקראי היא תופעת מעבר הפאזה. כלומר קיים ערך מסוים - , כך שכל עוד , הגרף לא מכיל מבנים מסוימים, וברגע ש- , מופיעים בגרף מבנים אלו. כפי שניתן לראות בתרשים 1, שנלקח מ- Albert & Barabasi (2002). ( Fig. 6 במאמר המקורי).

**תרשים 1 –** **תופעת מעבר הפאזה בגרפים אקראיים**



**תרשים 1:** הסתברויות הסף בהן מופיעים מבנים שונים בגרף אקראי. עבור הגרף מכיל קודקודים וקשתות מבודדים. עבור , מופיעים עצים בגודל 3. בעוד שעבור מופיעים כבר עצים בגודל 4. עבור קיימים כבר עצים מכל הסוגים, וכן מופיעים מעגלים מכל הסוגים. כאשר מופיעים תתי גרפים שלמים בגודל 4 (קליקות בגודל 4), ואילו עבור מופיעים תתי גרפים שלמים בגודל 5.

### מודל הגרף האקראי לעומת רשתות אמפיריות

מודל הגרף האקראי מייצר רשתות שהינן בעלות התכונה של "עולם קטן" ( Small World ), המאפיינת רשתות אמיתיות רבות. מספר השכנים במרחק מקודקוד מסוים שווה ל-, כאשר מניחים ש- קבוע. לכן על מנת לכלול את כלל קודקודי הרשת, הערך הנדרש עבור צריך לקיים , כאשר N הוא מספר הקודקודים ברשת. מכאן נובע כי המרחק האופייני בין זוג קודקודים בגרף אקראי עומד על דבר העונה על הדרישה להגדרת הרשת כבעלת תכונת "עולם קטן". אולם מלבד תכונה זו, כמעט כל שאר התכונות של מודל הגרף האקראי אינן תואמות לתכונות הקיימות ברשתות אמיתיות ( Newman, 2002 ): הואיל והסיכוי לקשר בין שני קודקודים הינו תמיד , ואינו תלוי בקיום של שכן משותף קודקודים מקדם האשכול שלו נמוך והינו ., אשר שואף ל-0 עבור ערכי . גדולים. בנוסף, כפי שכבר הוזכר למודל זה ישנה התפלגות דרגות פואסונית השונה מהתפלגויות דרגות הקודקודים המוכרות עבור רשתות אמפיריות. כמו-כן אין בו תיאום בין דרגות הקודקודים של קודקודים שכנים, אין בו מבנה קהילתי, הוא בעל שילוב תבניות אקראי לחלוטין. בקיצור כאשר מנסים ליצור מודלים של רשתות אמפיריות, מודל הגרף האקראי אינו בחירה טובה. עם כל זאת, חלק גדול מההבנה שלנו לגבי אופן הפעולה של רשתות נובע מהלימוד של מודל הגרף האקראי. במיוחד נושא מעבר הפאזה וקיום הרכיב הענק, אשר קיימים גם במודלים המתוחכמים יותר.

### הרחבות של מודל הגרף האקראי – מודל הקונפיגורציה

מודל הגרף האקראי ניתן להרחבה במספר דרכים על מנת ליצור באמצעותו רשתות הדומות יותר לרשתות האמפיריות. התכונה שקל ביותר לשלב במודל הינה התכונה של התפלגות דרגות קודקודים לא פואסונית, אשר מובילה למודל הקרוי "מודל הקונפיגורציה".

על מנת לייצר גרף בעל התפלגות קודקודים מסוימת, פועלים באופן הבא: ראשית מגדירים התפלגות דרגות קודקודים , כך ש- הינו אחוז הקודקודים ברשת שהינם בעלי דרגה . לכל אחד מהקודקודים בגרף משויכת דרגה סופית ע"פ ההתפלגות שבחרנו. בתחילה הקודקודים אינם מקושרים, כך שדרגתם הזמנית היא 0. כעת מתחילים לבחור באקראי זוגות של קודקודים שדרגתם הזמנית שונה מדרגתם הסופית, ומחברים אותם באמצעות קשת. חוזרים על תהליך זה עד אשר דרגת כל אחד מהקודקודים הינה הדרגה הסופית ששייכנו לו ע"פ ההתפלגות. תהליך זה מייצר בהסתברות שווה את כל הרשתות האפשריות בעלות התפלגות דרגות קודקודים שנבחרה. **מודל הקונפיגורציה** מוגדר כאוסף כל הגרפים שנוצרו באופן זה כאשר לכל אחד הסתברות שווה להיווצר.

מאז שנות ה-70 של המאה ה-20 נחקר מודל זה ע"י מספר רב של חוקרים (Bekessey et al,1972 , , Bender et al, 1978, Bollobas, 1980, Chung et al, 2002ועוד) וידועות עבורו תכונות רבות כגון: התנאי המדויק במונחי לקיומו של רכיב ענק, הגודל המדויק של הרכיב הענק, הכמות הממוצעת של קודקודים במרחק מסוים מקודקוד מרכזי, המרחק האופייני בין זוגות קודקודים ועוד.

## רשתות "עולם קטן"

מודל מתוחכם פחות אך קל יותר לניתוח ומעקב של רשת בעלת ערך מקדם Clustering גבוה הינו מודל ה"עולם קטן" שהוצע ע"י Watts & Strogatz (1998). המודל מתבסס על ההנחה שברשתות רבות יש מרכיב גיאוגרפי, אשר משפיע על קיום הקשרים בין הקודקודים ברשת. כלומר לקודקודים יש מיקום במרחב וסביר להניח כי קודקודים הקרובים זה לזה במרחב יהיו מקושרים זה לזה. מודל ה"עולם הקטן" מתחיל מנקודה זו בכך שבמצב ההתחלתי מסודרים קודקודי הרשת בשריג וקודקודים שמרחקם קטן ממרחק מסוים מקושרים ביניהם בקשת. תיאורטית ניתן לבנות רשתות ממודל זה עבור שריג ממימד כלשהו, אולם עיקר המחקר התמקד במקרה של שריג חד מימדי. במקרה זה, אנו מתחילים ברשת בה ישנם קודקודים המסודרים כשריג סגור חד מימדי, כלומר בצורת טבעת, ובה כל קודקוד מקושר לכל הקודקודים המרוחקים ממנו עד , ולכן בסך הכל ישנן קשתות. בשלב הבא מחווטים מחדש חלק מהקשתות הקיימות באופן אקראי. תהליך החיווט מחדש, מבוצע ע"י מעבר על כל אחת מהקשתות ובהסתברות מקושר הקצה שלה לקודקוד חדש שנבחר באקראי. ההגבלה היחידה על תהליך החיווט מחדש היא שהוא לא ייצור קשתות עצמיות או קשתות כפולות.

תהליך החיווט מחדש של הקשתות מאפשר למודל "העולם הקטן" לנוע בין המצב של שריג מסודר לבין מצב הדומה, אך לא זהה למצב של גרף אקראי. כאשר אנו מקבלים שריג מסודר. עבור מצב זה קל לראות כי מקדם האשכול הינו , אשר שואף ל- עבור ערכי גדולים. עם זאת, השריג המסודר אינו מציג את התכונה של "עולם קטן", כלומר המרחק הגיאודזי בין קודקודים שואף ל- עבור ערכי גדולים. כאשר כל קשת מחווטת מחדש למיקום חדש והרשת המתקבלת היא כמעט גרף אקראי בעל מרחק גיאודזי מסדר גודל של , אך בעלת מקדם אשכול מאוד נמוך . בין שני ערכים קיצוניים אלו קיימים טווחים נרחבים עבורם הרשת המתקבלת הינה בעלת מרחק גיאודזי נמוך וגם ערך מקדם Clustering גבוה. זוהי רשת "עולם קטן".

### תכונות המודל

התכונות של מודל "העולם הקטן" חושבו עבור מספר גרסאות של המודל. עבור הגרסה המקורית של Watts & Strogatz הוכח כי מקדם האשכול הינו . התפלגות דרגות הקודקודים במודל, אינה תואמת היטב את התפלגות דרגות הקודקודים המוכרת עבור רוב הרשתות האמפיריות. אולם נתון זה אינו מפתיע הואיל וזו לא הייתה מטרת המודל מלכתחילה. לכל קודקוד ישנם לפחות קשתות, כתוצאה ממבנה השריג ועוד קשתות הנובעות מתהליך החיווט מחדש אשר כמותן הינה בעלת התפלגות בינומית. ולכן ההסתברות לכך שלקודקוד יש דרגה שווה ל:

.

## מודלים בעלי מנגנון "קישור מועדף"

המודלים לעיל, מנסים לייצר רשתות המחקות תכונות ספציפיות של הרשתות הקיימות. אולם מודלים אלו אינם מסייעים להבין כיצד נוצרו תכונות אלו מלכתחילה ברשתות. קיימים מודלים שמטרתם העיקרית היא להסביר את התכונות של רשתות. מודלים אלו מבוססים על מנגנון האחראי על התפתחות הרשת כאשר הרשת בד"כ גדלה ע"י הוספה של קודקודים וצלעות באופן הדרגתי. מנגנון זה אמור לשקף את תהליך הגדילה שאולי קיים ברשת האמיתית המוביל ליצירת תכונות אלו. מחלקת הרשתות שנחקרה בצורה הכי מעמיקה באופן זה על מנת לגלות את תהליך היצירה שלה היא מחלקת הרשתות חסרות המידה ( Scale Free Networks ).

### המודל של Price

המודל של Price הינו אבטיפוס לרשתות בעלת התפלגות דרגות קודקודים מעריכית, המוכרות גם בשם רשתות חסרות מידה ( Scale Free ). Price חקר בשנות ה-60 רשתות ציטוטים של מאמרים מדעיים. רשתות אלו הוגדרו באופן הבא: כל מאמר מיוצג ע"י קודקוד ברשת. במידה ומאמר א' ציטט את מאמר ב' אזי הרשת הכילה קשת מכוונת מקודקוד א' לקודקוד ב'. Price (1965) גילה כי התפלגות דרגות היציאה והכניסה של הקודקודים ברשתות אלו (כמות המאמרים שצוטטו ע"י המאמר וכמות הציטוטים של המאמר במאמרים אחרים) מתפלגות באופן מעריכי. מספר שנים אח"כ פרסם Price מאמר נוסף (1976) ובו הציע הסבר, אשר משמש כיום כהסבר המקובל לתופעה. ההסבר מבוסס על רעיונות שפותחו ע"י Simon (1955), אשר הראה כי התפלגות מעריכית מתקבלת כאשר "העשירים נעשים עשירים יותר", כלומר כאשר הכמות שאתה מקבל הינה יחסית לכמות שכבר יש ברשותך. Price כינה תכונה זאת בשם "יתרון מצטבר" (Cumulative Advantage). כיום תכונה זו מוכרת יותר בשם "חיבור מועדף" (Preferential Attachment ) שם שנטבע ע"י Albert & Barabasi (1999).

תרומתו של Price הייתה החדרת הרעיון של Simon לתחום הרשתות ובאופן ספציפי לתחום של רשתות ציטוטים והתפלגות הדרגות שלהן. ההסבר שלו היה שהקצב בו מאמר מצוטט ע"י מאמרים חדשים הינו יחסי לכמות הציטוטים שכבר יש לו. כאשר ניתן להסביר זאת בקלות מבחינה הגיונית: ככל שמאמר מסוים מצוטט יותר כך גדל הסיכוי שיותר אנשים קראו אותו כאשר סקרו את הספרות הרלבנטית ולכן הסיכוי שהוא יצוטט גדל אף הוא. אותו הטיעון יכול לשמש גם עבור רשתות נוספות כמו רשת האינטרנט. Price אימץ, בעקבותיו של Simon, את ההנחה כי הקשר בין כמות הציטוטים הקיימת לבין ההסתברות להתווספות ציטוט חדש הינה ליניארית. עם זאת לא ברור אם הנחה זו אכן נכונה.

#### תיאור המודל של Price והפתרון שלו

נניח כי נתונה לנו רשת מכוונת בעלת קודקודים, למשל של רשת ציטוטים. נסמן ב- את אחוז הקודקודים ברשת שדרגת הכניסה שלהם היא , ולכן נקבל כי מתקיים . קודקודים חדשים מתווספים לרשת, אך לא בהכרח בקצב קבוע. לכל קודקוד שמתווסף יש דרגת יציאה מסוימת – כמות המאמרים שהוא מצטט אשר נקבעת בזמן יצירת הקודקוד. דרגת היציאה עשויה להשתנות בין קודקודים, אולם אנו מניחים כי דרגת היציאה הממוצעת, ברשת שנסמנה ב- היא קבועה לאורך זמן. היא גם דרגת הכניסה הממוצעת בגרף. ומתקיים . הואיל ודרגת היציאה עשויה להשתנות מקודקוד לקודקוד, עשוי לקבל ערכים לא שלמים, ואף ערכים קטנים מ-1.

בצורה הפשוטה ביותר של תהליך בעל "יתרון מצטבר", ההסתברות לקישור של קודקוד חדש לקודקוד קיים, כלומר ההסתברות שמאמר חדש יצטט מאמר קיים, היא יחסית לדרגת הכניסה של הקודקוד הקיים. עם זאת בתהליך זה ישנה בעיה הואיל וכל קודקוד מתחיל עם דרגת כניסה 0, ולכן ההסתברות שקודקודים חדשים יקשרו אליו את הצלעות שלהם תהיה 0. על מנת להתגבר על בעיה זו, הציע Price לשנות מעט את התהליך כך שההסתברות לקישור של קשת מקודקוד חדש לקודקוד קיים תהיה , כאשר הינו קבוע. במקרה של רשתות הציטוטים, השתמש Price ב- והצדיק זאת בכך שניתן להחשיב את הופעת המאמר כציטוט הראשון שלו. כלומר קיימת קשת עצמית בכל קודקוד חדש, ולכן ההסתברות לציטוט של מאמר קיים הינה . ההסתברות שקשת חדשה תקושר לקודקוד כלשהו בעל דרגת כניסה , היא:

הכמות הממוצעת המוספת של קשתות עבור כל קודקוד גם היא , ולכן הכמות הממוצעת של קשתות המקושרות לקודקודים בעלי דרגת כניסה היא . כמות הקודקודים בעלי דרגת כניסה , שהינה , קטנה בכמות זו, הואיל ולקודקודים אלו מתווספת קשת חדשה המגדילה את דרגת הכניסה שלהם ל-. עם זאת כמות הקודקודים בעלי דרגת כניסה , גדלה בגלל שישנם קודקודים שדרגת הכניסה שלהם גדלה מ- ל-. יוצאי הדופן היחידים הם קודקודים שדרגתם הייתה 0, ובהם הגידול שחל בדרגת הכניסה הינו בדיוק 1. אם מסמנים ב- את הערך בזמן שהרשת הכילה קודקודים, אזי השינוי שחל ב- עבור כל קודקוד שמתווסף שווה ל:

, *כאשר .*

*, כאשר* .

כאשר פותרים את מערכת המשוואות הזו מקבלים כי עבור ערכי  גדולים מתקיים: . כלומר, התפלגות דרגות הקודקודים הינה מעריכית עם מעריך .לכן בד"כ דרגות הקודקודים יתפלגו מעריכית עם מעריך בין 2 ל-3. (המעריך יהיה גדול מ-3 רק במקרה בו הינו בעל ערך קטן מ-1, דבר שלא סביר ברשת ציטוטים ). תחום ערכים זה למעריך של ההתפלגות תואם את הערכים שנצפו עבור רשתות אמיתיות נוספות**.** Price הראה כי תוצאות המודל שלו מתאימות לרשת הציטוטים המתקבלת מ- Science Citation Index.

נקודה נוספת הראויה לציון הינה כי במקרה של המודל של Price, המעריך המתקבל לא תלוי בערך ההתחלתי של . זאת בניגוד למודל של Albert & Barabasi המתואר בהמשך.

Price נתן רק פתרון אנליטי למודל שלו, מבלי לתת ערכים מספריים, הואיל ובתקופתו לא היו משאבים חישוביים שאפשרו ביצוע סימולציות על המודלים כפי שנעשה ע"י חוקרים בשנים האחרונות.

בשנים האחרונות נעשתה התקדמות רבה בהבנת תהליכים בעלי "חיבור מועדף" וגידול רשתות. עם זאת רוב העבודה שנעשתה התבססה על מודל מעט שונה, שהינו המודל של Albert & Barabasi המתואר להלן.

### המודל של Albert & Barabasi

המנגנון של "יתרון מצטבר" שהוצע ע"י Price מקובל כיום כהסבר להתפלגות מעריכית של דרגות הקודקודים שנצפתה לא רק ברשתות ציטוטים אלא גם ברשתות רבות אחרות במגוון תחומים. עם זאת עבודתו של Price פחות מוכרת בקהילה המדעית, ולכן מנגנון זה לא צבר תאוצה עד לגילויו מחדש מספר עשורים אח"כ ע"י Albert & Barabasi אשר נתנו לו שם חדש "חיבור מועדף" ( Preferential attachment ). במאמר רב השפעה שהם פרסמו ב-Science (Albert & Barabasi, 1999) הם הציעו מודל גידול לרשת האינטרנט שהינו מאוד דומה למודל של Price אך עם הבדל חשוב אחד.

המודל של Albert & Barabasi זהה למודל של Price בכך שהקודקודים מוספים לרשת עם דרגה , שאינה משתנה אח"כ והקצה השני של הקשת מתחבר לקודקוד קיים בהסתברות פרופורציונאלית לדרגת הקודקוד. ההבדל בין שני המודלים הינו בכך שהקשתות במודל של Albert & Barabasi הינן לא מכוונות, ולכן אין הבחנה בן דרגת הכניסה לדרגת היציאה של הקודקוד. לדבר זה יש יתרונות וחסרונות. מחד הן רשתות ציטוטים והן רשת האינטרנט הינן במציאות רשתות מכוונות, לכן כל מודל שהינו לא מכוון מחמיץ מאפיין מהותי של הרשת. מאידך, ההתעלמות מהכיווניות של הרשת, מאפשר למודל לעקוף את הבעיה במודל של Price – איך הקודקוד החדש מקבל את הקישור הראשון. כל קודקוד חדש מתחיל עם דרגה , ולכן יש לו באופן אוטומטי הסתברות גדולה מ-0 לכך שיקושרו אליו קשתות חדשות. כך שניתן לומר שמודל זה מקריב חלק מהיכולת לתאר את הרשת באופן קרוב למציאות למען פישוט המודל. התוצאה העיקרית של הבדל זה הינה בכך שעבור המודל של Albert & Barabasi התפלגות דרגות הקודקודים ברשת היא מעריכית עם ערך יחיד . אולם הכללות שבוצעו בהמשך למודל טיפלו בנקודה זו. עבור המודל ההסתברות שקשת חדשה תקושר לקודקוד בעל דרגה היא: *.* הערך במכנה שווה לדרגת הקודקודים הממוצעת ברשת שהינה . דבר זה נובע מכך שלכל קודקוד ישנם קשתות, והואיל והרשת איננה מכוונת, כל הוספה של קודקוד תורמת לכמות הקשתות ברשת. כמות הקודקודים בעלי דרגה אליהם מקושרת קשת נוספת כאשר מוסיפים קודקוד חדש לרשת הינה , אשר אינה תלויה ב-. לכן כמות הקודקודים בעלי דרגה קטנה בכמות זו. הואיל וקודקודים אלו הופכים להיות בעלי דרגה . מצד שני כמות הקודקודים בעלי דרגה גדלה בגלל הקודקודים בעלי דרגה אליהם התווספה קשת. אם מסמנים ב- את הערך בזמן שהרשת הכילה קודקודים, אזי השינוי שחל ב- עבור כל קודקוד שמתווסף הינו:

כאשר

ו- כאשר ואין קודקודים בעלי דרגה .

כאשר פותרים את מערכת המשוואות הזו מקבלים כי עבור ערכי גדולים מתקיים: .

מלבד הפתרון הבסיסי עבור התפלגות דרגות הקודקודים, ידועות תוצאות רבות נוספות עבור מודל זה.

בניתוח מעמיק שעשו למודל הראו Krapivsky & Redner (2001) כי ישנה התאמה בין "גיל" הקודקוד לבין הדרגה שלו, כאשר קודקודים ותיקים יותר הינם בעלי דרגה ממוצעת גבוהה יותר, ולכן התפלגות דרגות הקודקודים של הגרף נובעת בעיקרה כתוצאה מהוותק של הקודקודים. Adamic & Hubberman (2000) הראו באמצעות בדיקת מידע מרשת ה-WWW כי אין הדבר נכון, והסבירו זאת בכך שישנה חשיבות למידת המשיכה של הקודקוד ( האתר ), כאשר קודקודים בעלי משיכה גבוהה יהיו בעלי דרגה גבוהה יותר מאשר קודקודים ותיקים יותר בעלי משיכה נמוכה. מתוצאה זו נובע כי המודל אינו מדמה לגמרי את רשת ה-WWW. עם זאת אין בכך סתירה לכך שהתפלגות הקודקודים ברשת האינטרנט נובעת כתוצאה ממנגנון המיישם "חיבור מועדף".

בשני המודלים, של Price ושל Albert & Barabasi מניחים כי יש שימוש ב"חיבור מועדף" ליניארי. מספר מחקרים ( Newman, 2001 ,Jeong et al, 2003 ) בדקו את קצב השינוי בדרגת הקודקודים לאורך זמן של רשתות שונות כגון: רשת האינטרנט, רשתות ציטוטים ורשתות שיתופי פעולה בין שחקנים. תוצאות הבדיקה הראו כי השינוי בדרגת הקודקודים היה די פרופורציונאלי לדרגה ההתחלתית שלהם ולכן הגיעו למסקנה כי על הרשת עבר תהליך המבוסס על "חיבור מועדף ליניארי".

#### הרחבות של המודל של Albert & Barabasi

המודל של Albert & Barabasi זכה להתעניינות רבה. התעניינות זו באה לידי ביטוי בשני מישורים: ניתוח של המודל עצמו והצעת מודלים רבים המכלילים את הרעיון. מטרת ההכללות הייתה בין השאר לאפשר יצירת רשת בעלת התפלגות דרגות קודקודים מעריכית עם חזקה השונה מ-3. סקירה מעמיקה של תחום הרשתות שפורסמה ע"י Albert & Barabasi (2002) מכילה בין השאר סיכום ממצה של ההכללות ( ראה טבלה III במאמר הנ"ל ).

באופן כללי ההכללות למודל מתייחסות למספר היבטים:

* שינוי פונקצית ההסתברות לקישור בין קודקוד חדש לקודקוד קיים, כך שבמקום שהיא תהיה מבוססת באופן ליניארי על דרגת הקודקוד בלבד, היא תהיה מבוססת למשל על עבור קבוע כלשהו, או שתהיה מבוססת על דרגת הקודקוד אך לא באופן ליניארי, כלומר , עבור קבוע כלשהו.
* שינוי דרגת הקודקודים הממוצעת לאורך הזמן: ישנם עדויות לכך כי דרגת הקודקודים הממוצעת ברשת ה-WWW איננה קבועה אלא הולכת וגדלה לאורך זמן. Dorogovtsev & Mendes (2001) הציעו הכללה של המודל המשנה את ערכו של - כמות הקשתות שקודקוד חדש מקשר לקודקודים קיימים - לאורך זמן. השינוי בערכו של לאורך הזמן מבוצע בעקיפין ע"י התחשבות בגודל הרשת. כלומר גדל בהתאמה לגודל הרשת כפונקציה של , עבור קבוע כלשהו. ההסתברות לקישור קשת חדשה לקודקוד קיים פרופורציונאלית ל- עבור קבוע. הם הראו כי עבור מודל זה מתקבלות רשתות עם התפלגות דרגות קודקודים מעריכית עם מעריך .
* יצירה דינמי של קשתות: המודל של Albert & Barabasi והמודל של Price מנסים לתאר את התפתחות של רשתות. בשני מודלים אלו, קיומן של הקשתות נקבע ברגע יצירת הקודקוד החדש ולא משתנה אח"כ. אולם בעוד עבור רשתות ציטוטים ההנחה כי הקשתות לא משתנות היא הגיונית, ברשת ה-WWW הדבר אינו נכון. קישורים חדשים מתווספים לדפים קיימים וקישורים קיימים נמחקים באופן דינמי. מספר מודלים התומכים ביכולת זו הוצעו (Dorogovtsev & Mendes, 2000 , Tadi'c, 2001), כאשר חלקם תומכים בהוספת קשתות חדשות ואילו חלקם מאפשרים גם מחיקת קשתות קיימות.
* התחשבות בגורמי משיכה נוספים: כפי שתואר לעיל, במודל של Albert & Barabasi ישנה התאמה בין "גיל" הקודקוד לבין הדרגה שלו, בעוד שברשת ה-WWW, הדבר לא בהכרח נכון. על מנת לתמוך בכך, הוצעו מודלים (Bianconi & Barabasi, 2001) המעניקים לכל קודקוד רמת "משיכה" הנבחרת מתוך התפלגות מסוימת , כך שההסתברות לקישור קשת חדשה לקודקוד קיים היא פונקציה של דרגת הקודקוד ורמת המשיכה שלו .

#### מודלי גידול נוספים לרשתות

המודל של Albert & Barabasi, כמודל לתיאור רשת ה-WWW, הינו בעל מספר חסרונות, וביניהם:

* המודל יוצר רשת לא מכוונת, בעוד ה-WWW הינה רשת מכוונת.
* הרשת הנוצרת הינה בעלת רכיב קשירות אחד, לעומת רשת ה-WWW המכילה מספר רכיבי קשירות רבים.
* במידה ומשנים את המודל כך שיהיה מכוון, אזי מקבלים רשתות חסרות מעגלים, עם רכיבי קשירות חלשים.
* התפלגות דרגות היציאה ברשת ה-WWW היא מעריכית, ואילו עבור המודל דרגת היציאה היא קבועה.

רוב החסרונות נכונים גם עבור המודל של Price, אולם המודל של Price יועד מלכתחילה עבור רשתות ציטוטים. הואיל ועבור תחום זה, חסרונות אלו אינם מהווים בעיה ולכן המודל דווקא הגיוני.

על מנת להתמודד עם בעיות אלו הוצעו מספר מודלים חליפיים.:

1. Callaway et al (2001), הציע מודל שמטרתו להתגבר על הבעיה של קיום רכיב קשירות בודד. המודל המוצע זהה למודל של Price, אך עם הבדל משמעותי אחד – בכל שלב מתווסף קודקוד חדש, אך בניגוד למודל של Price, הקשתות לא מקושרות לקודקוד החדש אלא מקושרות לקודקודים קיימים הנבחרים באקראי. מודל זה מאפשר ליצור רשתות בעלות מספר רכיבי קשירות (בהתאם לערך של ). למרות שמבחינה שטחית נראה כי מודל זה זהה למודל הגרף האקראי, בדיקה מעמיקה של המודל מגלה כי אין הדבר כך, במודל זה הקודקודים הותיקים נוטים להיות מקושרים אחד לשני. לרשתות הנוצרות באמצעות מודל זה אין התפלגות דרגות מעריכית, אולם הוא הוכלל על מנת לתמוך ב"קישור מועדף" ע"י Dorogovtsev et al (2001), ע"י קישור הקשתות לקודקודים הנבחרים בהתאמה לדרגות הקודקודים ועוד קבוע, על מנת לתת סיכוי גם לקודקודים בודדים להיות מקושרים.
2. Krapivsky & Redner (2002) הציעו מודל אחר. במודל זה בכל שלב מתווסף קודקוד חדש או קשת מכוונת המקשרת בין זוג קודקודים הנבחרים ע"פ דרגות הכניסה והיציאה שלהם ועוד היסט קבוע. מודל זה מאפשר הוספת קשתות לאחר יצירת הקודקודים ומייצר רשתות בעלות התפלגות דרגות קודקודים מעריכית הן בדרגות הכניסה והן בדרגות היציאה. ע"י שינוי ההיסטים ניתן אף לשלוט במעריך של התפלגות דרגות הקודקודים השונות. כמו-כן מודל זה מאפשר קיום של מספר רכיבי קשירות. מודל זה פותח על בסיס של מודל קודם, אשר תמך בכל האפשרויות לעיל, אולם יצר רכיב קשירות בודד, בכך שכל קודקוד חדש שנוצר קושר מיד לאחד הקודקודים ברשת. (שני מודלים אלו הינם חלק מהמודלים שנבחנו במסגרת המחקר, כפי שמתואר בהמשך העבודה ).

## מודלים להעתקה ושינוי של קודקודים

ישנן רשתות אשר מציגות התפלגות דרגות מעריכית, אולם מעריכים כי הן לא נוצרו כתוצאה של שימוש במנגנון של "קישור מועדף" מהסוג שהוצע ע"י Albert & Barabasi (1999). דוגמא לכך הן רשתות ביוכימיות מסוגים שונים, כגון רשתות המייצגות את האינטראקציות בין חלבונים'. רשתות אלו משתנות על פני טווחי זמן ארוכים מאוד כתוצאה משינויים אבולוציוניים, ומניחים כי התהליך באמצעותו הן נוצרו מורכב יותר מאשר תהליך מסוג "קישור מועדף" פשוט. מספר הסברים אפשריים ליצירת רשתות אלו הוצעו:

### מנגנון קישור לקשתות

אחד המנגנונים הפשוטים ביותר, ככל הנראה, ליצירת רשתות בעלות התפלגות דרגות מעריכית, אשר אינו משתמש באופן מפורש ב"קישור מועדף" הוצע ע"י Dorogovtsev et al (2001a). במודל המוצע בכל שלב בוחרים באקראי את אחת הקשתות הקיימות ברשת. לאחר בחירת הקשת, מוסיפים לרשת קודקוד חדש ומקשרים אותו לשני הקודקודים אליהם הייתה מקושרת הקשת שנבחרה. כתוצאה מהתהליך זה ההסתברות שקשת תקושר לקודקוד מסוים פרופורציונאלית לדרגה שלו. במילים אחרות הרשתות הנוצרות באמצעות מודל זה הינן בעלות אותה צורה "קישור מועדף" כמו המודל של Albert & Barabasi , אשר בין השאר באה לידי ביטוי בכך שהתפלגות הדרגות שלהן דומה.

### מנגנון העתקת קודקודים

Kleinberg et al (1999) הציע הסבר אפשרי ליצירת רשת ה-WWW באמצעות מנגנון "העתקה של קודקודים". המודל המוצע מבוסס על ההנחה שעמוד חדש ברשת ה-WWW העוסק בנושא מסוים יעתיק קישורים מעמודים קיימים העוסקים בנושא. בניית הרשתות באמצעות המודל מבוצעת על ידי האלגוריתם הבא: בכל שלב מוסיפים קודקוד חדש לרשת ובוחרים מספר של קשתות שיקושרו ממנו לקודקודים אחרים. היעד של הקשתות נבחר באופן הבא: בוחרים באקראי קודקוד קיים ברשת אשר ישמש כאבטיפוס. עבור כל קשת בקודקוד האבטיפוס האלגוריתם מחליט בהסתברות , האם ליצור קשת מכוונת מהקודקוד החדש לקודקוד אקראי ברשת, או ( בהסתברות ) להעתיק את הקשת וליצור קשת מכוונת מהקודקוד החדש ליעד של הקשת. ממשיכים בתהליך זה עד אשר דרגת היציאה של הקודקוד החדש היא . במידה ולקודקוד שנבחר באקראי לא היו מקושרות קשתות, חוזרים על תהליך זה עם קודקוד נוסף שנבחר באקראי וכך הלאה עד שמקשרים את כלל הקשתות. בצורה הכללית ביותר שלו המודל של Kleinberg et al כולל גם מנגנונים למחיקה של קשתות וקודקודים.

מנגנון זה מוביל ליצירת רשתות בעלות התפלגות דרגות מעריכית מהסיבה הבאה: הסיכוי שקשת אקראית מקודקוד תוביל לקודקוד בעל דרגת כניסה פרופורציונאלי ל-, ולכן הקצב בו גדלה דרגת הקודקוד פרופורציונאלי לדרגה הנוכחית שלו. כמו במודל של Price מודל זה לא יוסיף קשתות לקודקודים שדרגתם 0, ולכן נכללת במודל הסתברות קבועה לכך שהיעד של קשת חדשה יהיה קודקוד עם דרגה 0. במאמר המשך (Kumar et al, 2000) הוכח כי מודל זה מייצר רשתות בעלות התפלגות דרגות קודקודים מעריכית בעלת מעריך השווה ל-, כאשר הינו היחס בין מספר הקשתות שהיעד שלהן נבחר באקראי לעומת מספר הקשתות שהיעד שלהן נבחר כתוצאה מהעתקה. עבור ערכים קטנים של מתקבלות רשתות בעלות התפלגות דרגות קודקודים מעריכית בעלות מעריך שהינו הטווח שנצפה עבור רוב הרשתות האמפיריות. לא ברור אם מנגנון ההעתקה הינו הגורם להתפלגות דרגות הקודקודים ברשת ה-WWW, אולם מודל זה עורר התעניינות רבה בהקשר של רשתות ביולוגיות, הואיל ובמקרה זה נראה הגיוני כי גנים חדשים עשויים להיווצר עקב שכפול של גנים אחרים יותר מפעם אחת בטעות. לגנים המשוכפלים יהיו אותו התפקוד, דבר השקול להעתקת הקישורים במודל., אולם לאורך זמן, ייתכן כי ייווצר הבדל בין גנים אלו כתוצאה מתהליכים התפתחותיים.

### מנגנון הליכה על הרשת

מנגנון ההליכה על הרשת הוצע ע"י Vasquez (2002). המנגנון הוצע מתוך ניסיון לחיקוי התהליך המתרחש ברשתות ציטוטים. באופן כללי, כאשר מתכננים לכתוב מאמר בתחום חדש, לכותב ישנה היכרות רק עם מספר מצומצם של מאמרים חשובים. מציאת מאמרים רלבנטיים נוספים מבוצעת באמצעות שימוש בהפניות המופיעות במאמרים אלו. תהליך זה נמשך באופן רקורסיבי, ובסופו של דבר המאמר הסופי יכיל הפניות שהתגלו באמצעות תהליך זה. Vasquez ניסח אלגוריתם ליצירת רשתות המחקה את התהליך, אשר פועל באופן הבא: מתחילים עם רשת בעלת קודקוד בודד. בכל שלב מוסיפים קודקוד חדש אשר מקושר לקודקוד קיים אשר נבחר באופן אקראי. קודקוד זה מתווסף לרשימת קודקודים למעקב. עבור כל קודקוד ברשימת המעקב מחליטים עבור כל אחד מהקודקודים אליהם הוא מקושר בהסתברות האם ליצור קשת מכוונת מהקודקוד החדש לקודקוד זה. קודקודים אשר נוצר אליהם קשר מתווספים גם הם לרשימת המעקב. כאשר מסיימים לעבור על כל הקשרים של קודקוד מסוים ברשימת המעקב הוא מוסר ממנה. תהליך זה נמשך עד אשר אין קודקודים נוספים לעבור עליהם. במקרה בו , ניתן לראות כי קודקודים בעלי דרגה גבוהה הינם בעלי הסתברות גבוהה יותר לכך שייווצר אליהם קישור חדש. כתוצאה מכך מתקבלות רשתות בעלות התפלגות דרגות קודקודים מעריכית בעלות מעריך . כאשר משתנה בין 0 ל-1, הראה Vasquez, באמצעות סימולציות שעשה, כי ישנו מעבר פאזה: עבור התפלגות דרגות הקודקודים דועכת בצורה אקספוננציאלית, ואילו עבור  *היא מעריכית עם מעריך בעל ערך קרוב מאוד ל-2 שהינו הערך המתקבל* במקרה בו . כך שבעוד שהמודל לא כולל באופן מפורש "קישור מועדף" המנגנון האחראי על יצירת הקשתות יוצר זאת.

# מטרת המחקר

מטרתו של מחקר זה הייתה לבדוק האם קבוצה של רשתות לימוד מקוונות אמפיריות ניתנת לסיווג למספר מצומצם של מודלים ( אולי אף מודל יחיד ). זאת אומרת שהמבנה של הרשתות אשר סווגו למודל מסוים, מתאים למודל זה באופן **ניכר משמעותית** יותר מאשר לשאר המודלים. (רמת הדיוק של ההתאמה מוגדרת בחלק המתודולוגיה ) הואיל וישנם מודלים רבים מאוד ליצירת רשתות, ולא ניתן לבדוק את כולם, נבחרו מספר מודלים מצומצם המייצגים מנגנונים שונים ליצירת רשתות. בחינת הרשתות בוצעה בשני אופנים, אשר בחנו את נתוני הרשתות משני היבטים:

* **בחינת קישוריות הרשתות – Binary Case:** בחינה זו מתבססת על מבנה הקישוריות ברשתות, בלבד. בבדיקה זו לא נעשה שימוש במידע על עוצמת הקשרים בין הקודקודים ברשת. זוג קודקודים נחשב למקושר במידה וקיים ביניהם קשר כלשהו, כלומר עוצמת הקשר בין זוג הקודקודים הינה לפחות 1.
* **בחינת עוצמת הקשרים – Weighted Case :** זוהי בדיקה מקיפה יותר של נתוני הרשתות, אשר מתבססת הן על מבנה הקישוריות ברשתות והן על עוצמתם של הקשרים. רעיון זה הוזכר בעבודתם של Middendorf et al (2004), אולם לא מומש עד כה

**התועלות הצפויות של מחקר זה הן:**

1. זיהוי המודל המתאים ליצירת רשתות יוביל להבנה טובה יותר של התנהגות המשתמשים ברשתות לימוד מקוונות. בכך יאפשר מחקר זה הבנה טובה יותר של התהליכים המתרחשים במהלך יצירת ידע משותף.
2. תוצאת המחקר יכולה לכוון את מנחה רשתות הלימוד המקוונות בתכנון המשתתפים ברשת ותפקידיהם.
3. המחקר יספק כלי תוכנה לזיהוי המנגנונים השולטים בהתנהגות רשתות לימוד מקוונות על ציר הזמן.

# בסיס המידע

האוניברסיטה הפתוחה ישראל הינה מוסד ללימוד מרוחק, אשר הלימודים בו מבוססים באופן משמעותי על טכנולוגיות לימודיות. אופי הלימוד באוניברסיטה הינו עצמאי בעיקרו, כאשר בחלק מהמקרים ישנם מפגשים עם מנחה מטעם האוניברסיטה . המפגשים מתמקדים בהיבטים שונים של הקורס, כגון: העברת חומר לימודי ופתרון בעיות לדוגמא. בנוסף למפגשים, בכל קורס ישנו שימוש ברשת לימוד מקוונת, אשר פעילה בד"כ לאורך סמסטר בודד ( כ-16 שבועות ). רשת זו משמשת כפורום של הסטודנטים וכן של מנחה הקורס. המטרות של הרשת יכולות להיות: החל מבניית ידע שיתופי בין הסטודנטים, העברת הודעות מהמנחה לסטודנטים בנוגע לקורס, מענה של המנחה על שאלות הסטודנטים, תמיכה טכנית, ואף מטרות חברתיות. אותה רשת לימוד מקוונת יכולה לשרת יותר ממטרה אחת במקביל. גודל הרשת, קשרי התגובות ותבנית ההשתתפות משתנים מקורס לקורס ומסמסטר לסמסטר. כמות המשתתפים ברשת נעה בין 10 ל-150, אולם ברובן מדובר על כ-50 סטודנטים. בפורום ניתן לכתוב הודעות, אשר ניתנות לצפייה ע"י כלל המשתתפים בקורס.

במחקר זה נבחנו 35 רשתות לימוד מקוונות מתוך כ-500 רשתות של האוניברסיטה הפתוחה בישראל. הרשתות נבחרו באקראי, כאשר לא נכללו בבחירה 5 רשתות אשר כמות המשתתפים בהם הייתה נמוכה מהסף שנבחר של 10 משתתפים. הייצוג הרשתי של רשתות הלימוד שנבחנו, נבנה על סמך בסיס נתונים המכיל את ההודעות שנותח באמצעות תוכנה בשם opus2sna, אשר פותחה במסגרת מחקרים קודמים אשר נערכו על נתוני הרשתות אלו (Aviv et al., 2007a, Aviv et al., 2007b ). הרשתות הנוצרות באמצעות התוכנה בנויות באופן הבא: כל משתתף מיוצג ע"י קודקוד ברשת. קשת מכוונת מקודקוד לקודקוד נוצרת במידה והמשתתף המיוצג ע"י קודקוד הגיב להודעה של המשתתף המיוצג ע"י קודקוד . מספר התגובות בין זוג משתתפים מגדיר את עוצמת הקשר ביניהם.

# מודלים ליצירת רשתות

כפי שצוין במטרת המחקר, נערכו שתי בדיקות של נתוני הרשתות: בדיקת הקישוריות ובדיקת עוצמת הקשרים. עבור כל אחת מהבדיקות נבדקה קבוצה נפרדת של מודלים. באופן כללי המודלים שנבדקו במקרה של בדיקת עוצמת הקשרים הינם הגרסה הממושקלת (Weighted) של המודלים שנבדקו בבדיקת הקישוריות.



## מודלים לבחינת קישוריות הרשתות

כפי שתואר בסקירה על תחום הרשתות בתחילת עבודה זו, בשנים האחרונות הוצע מגוון מודלים נרחב ליצירת רשתות המתבססים על מנגנונים שונים. מתוך מגוון זה נבחרו שישה מודלים שונים המייצגים מנגנונים שונים ליצירת רשתות אקראיות מכוונות והם:

### מודל הגרף האקראי המכוון

זוהי הגרסה המכוונת של המודל הקלאסי ליצירת גרפים אקראיים **(**Erdos & R'enyi, 1960 **)**, אשר תואר בהרחבה בסקירה בסעיף ‎2.3.1. מודל זה מאופיין בכך שהקשרים בין הקודקודים ברשת נבחרים באופן אקראי לחלוטין, מבלי שישנו מנגנון כלשהו הנותן עדיפות ליצירת קשרים מסוימים על פני קשרים אחרים.

**הפרמטרים של המודל:**

* + N - כמות הקודקודים ברשת.
  + - פרמטר המשמש לקביעה האם לייצר קשתות בין זוג קודקודים.

**תהליך בניית הרשת באמצעות המודל מבוצע באופן הבא:**

* + קובעים את כמות הקודקודים ברשת N, ואת ערכו של הפרמטר בטווח .
  + עבור כל זוג קודקודים **שונים** ברשת מחליטים בהסתברות האם ליצור קשת מכוונת מקודקוד לקודקוד .

**תהליך זה מוביל ליצירת רשתות בעלות התכונות הבאות:**

* מספר הקשתות הממוצע ברשת הינו .
* המרחק הממוצע בין זוגות הקודקודים גדל כפונקציה של .
* התפלגות דרגות הקודקודים בגרף היא בינומית.

**אופן בחירת הפרמטרים עבור המודל**

על מנת ליצור רשתות הדומות לרשתות האמפיריות נבחר את להיות שווה למספר הקודקודים ברשת האמפירית ואילו ערכו של יהיה , כאשר הינו מספר הקשתות ברשת האמפירית.

### מודל "קישור מועדף" סטטי

מודל זה מייצג את משפחת המודלים הסטטיים בעלי תכונת ה"קישור המועדף". המודל הוצע לתיאור התכונות של רשתות גדולות רבות (Goh, Khang, & Kim, 2001 ). המודל מבוסס על יצירת רשתות סטטיות, בהן כמות הקודקודים ברשת נקבעת בזמן יצירת הרשת. קודקודים אלו מקבלים רמות משיכה שונות באופן אפריורי המתפרסות על טווח ערכים נרחב. במהלך בניית הרשת, הקשרים "מעדיפים" להיווצר בין קודקודים בעלי רמות משיכה גבוהות.

**הפרמטרים של המודל:**

* + – כמות הקודקודים ברשת.
  + – פרמטר לקביעת מספר הקשתות.
  + - פרמטר השולט על תהליך יצירת הקשתות ברשת.
  + - פרמטר השולט על תהליך יצירת הקשתות ברשת.

**תהליך בניית הרשת באמצעות המודל מבוצע באופן הבא:**

* קודקודי הרשת ממוספרים מ-1 עד .
* לכל קודקוד משויכים שני משקלים:
  + משקל נכנס, כאשר המשקל הנכנס של קודקוד הינו .
  + משקל יוצא, כאשר המשקל היוצא של קודקוד הינו .
* בוחרים באקראי שני קודקודים שונים.
  + קודקוד מקור, כאשר ההסתברות לבחור את הקודקוד שמספרו היא בהסתברות .
  + קודקוד יעד, כאשר ההסתברות לבחור את הקודקוד שמספרו היא בהסתברות .
* במידה ולא קיימת ברשת קשת מכוונת בין הקודקודים שנבחרו - מקודקוד לקודקוד - אזי מוסיפים לרשת קשת זו.
* חוזרים על התהליך עד אשר כמות הקשתות ברשת מגיעה ל-.

**תהליך זה מוביל ליצירת רשתות בעלות התכונות הבאות:**

* הדרגה הנכנסת הממוצעת של קודקוד ברשת היא .
* הדרגה היוצאת הממוצעת של קודקוד ברשת היא .
* - הדרגה הנכנסת של קודקוד מקיימת:   
   , כאשר .
* - הדרגה היוצאת של קודקוד מקיימת:  
   ,כאשר .

מכך נובע כי התפלגות דרגות הכניסה של הקודקודים היא בעלת התפלגות מעריכית מהצורה הבאה: . כאשר . לכן ע"י התאמת הפרמטר בטווח ניתן לקבל ערכים שונים של הפרמטר בטווח .

כמו-כן התפלגות דרגות היציאה של הקודקודים היא בעלת התפלגות מעריכית מהצורה הבאה: . כאשר . לכן ע"י התאמת הפרמטר בטווח ניתן לקבל ערכים שונים של הפרמטר בטווח .

**אופן בחירת הפרמטרים עבור המודל**

על מנת ליצור באמצעות המודל רשתות דומות לרשתות האמפיריות, נבחר את להיות שווה למספר הקודקודים ברשת האמפירית ואילו הפרמטר ייבחר כך שכמות הקשתות תהיה שווה לערך השלם הקרוב ביותר של כמות הקשתות ברשת האמפירית.

תיאורטית, ניתן לחשב במדויק את ערכי הפרמטרים ו- ע"פ התפלגות דרגות הקודקודים ברשת האמפירית. אולם באופן מעשי הדבר איננו אפשרי עקב כמות הקודקודים הנמוכה יחסית ברשתות האמפיריות, אשר גורמת לכך שאין ברשותנו מספיק מידע לזיהוי מדויק של התפלגות דרגות הקודקודים. לכן במסגרת הבדיקה בחרנו באופן שרירותי את הערך 0.5 עבור פרמטרים אלו.

### מודל "קישור מועדף" דינמי

מודל זה מייצג את משפחת המודלים המתפתחים בעלי תכונת ה"קישור המועדף". מודל זה הוצע לתיאור ההתפתחות של ה-World Wide Web (Krapivsky, Rodgers, & Redner, 2001 ). זהו מודל התפתחותי בו מוסיפים באופן הדרגתי לרשת קודקודם וקשתות.

**הפרמטרים של המודל:**

* + – כמות הקודקודים ברשת.
  + – פרמטר השולט על ההסתברות להוספת קודקוד או קשת לרשת.
  + - פרמטר המשפיע על בחירת קודקוד הכניסה של הקשת שרוצים להוסיף.
  + - פרמטר המשפיע על בחירת קודקוד היציאה של הקשת שרוצים להוסיף.

**תהליך בניית הרשת באמצעות המודל מבוצע באופן הבא:**

* + בהסתברות מוסיפים לרשת קודקוד חדש. קודקוד זה מקושר לאחד הקודקודים הקיימים כאשר הקודקוד אליו הוא מקושר נבחר באקראי ע"פ דרגות הכניסה של הקודקודים הקיימים. כך שההסתברות לקשר את הקודקוד החדש לקודקוד היא , כאשר הינה דרגת הכניסה של קודקוד .
  + בהסתברות מוסיפים לרשת קשת בין שני קודקודים קיימים כאשר   
    הקודקודים נבחרים באקראי ע"פ דרגות הכניסה והיציאה שלהם. ההסתברות לבחור את הקודקוד כקודקוד הכניסה של הקשת החדשה היא , כאשר הינה דרגת הכניסה של קודקוד . ההסתברות לבחור את הקודקוד כקודקוד היציאה של הקשת החדשה היא +b, כאשר הינה דרגת היציאה של קודקוד .

תהליך בניית הרשת מתמשך עד אשר מספרי הקודקודים והקשתות שהרשת מכילה מתאימים לערכיהם ברשת האמפירית.

תהליך זה מייצר רשתות אשר הדרגה הממוצעת של קודקודים בהם היא .

הואיל ותהליך בניית הרשת לא מאפשר לשלוט במדויק במספר הקשתות הנוצרות, נעצור את התהליך כאשר מספר הקודקודים ישתווה לכמות הקודקודים ברשת האמפירית. בנוסף, על מנת להגדיל את הדמיון לרשת האמפירית נבחר את כך ש- יהיה שווה לדרגה הממוצעת ברשת האמפירית. כתוצאה מכך נקבל כי מתקיים: ולכן *.*

* *במסגרת עבודה זו קבענו את ערכם של הפרמטרים ו- ל-.*

### מודל "קישור מועדף" דינמי משופר

מודל זה הינו גרסה מאוחרת יותר של מודל מס' 3, אשר פותחה ע"י הממציאים של המודל המקורי ( Krapivsky, & Redner, 2002 ). מודל זה פותח ממודל מספר 3, על מנת לאפשר יצירת רשתות דומות יותר לרשתות הקיימות במציאות. ההבדל בין המודלים הינו בכך, שהרשתות הנוצרות באמצעות מודל זה עשויות להכיל מספר רכיבי קשירות, בעוד שהרשתות המיוצרות באמצעות מודל מס' 3 מכילות רכיב קשירות בודד. הבדל זה נובע מכך שבמודל זה קודקודים חדשים לא מקושרים לקודקודים קיימים מייד עם היווצרותם.

**הפרמטרים של המודל:**

* + – כמות הקודקודים ברשת.
  + – כמות הקשתות ברשת.
  + - פרמטר המשפיע על בחירת קודקוד הכניסה של הקשת שרוצים להוסיף.
  + - פרמטר המשפיע על בחירת קודקוד היציאה של הקשת שרוצים להוסיף.

בניית הרשת מבוצעת בדומה לתהליך בניית הרשת במודל מס' 3. ההבדל היחיד הינו בכך שבמקרה בו נבחרה האפשרות להוסיף קודקוד חדש, הוא אינו מקושר לקודקוד קיים באופן אוטומטי.

באופן מעשי, בניית הרשת בוצעה באופן שונה במקצת מאשר בוצע עבור המודל המקורי. התהליך שבוצע בו שימוש בפועל הוגדר באופן הבא:

**שלב א'**

יוצרים רשת בעלת N קודקודים ללא צלעות.

**שלב ב'**

מוסיפים לרשת קשתות. בכל פעם שמוסיפים קשת בוחרים באקראי את הקודקודים שהיא תקשר ע"פ דרגות הכניסה והיציאה שלהם. ההסתברות לבחור את קודקודי היציאה והכניסה של הקשת הינן כפי שהוגדרו עבור מקרה הוספת הקשת במודל המקורי.

חוזרים על תהליך הוספת הקשתות עד אשר ישנם קשתות ברשת. כאשר **חשוב להדגיש** כי הקודקודים ביניהם תהיה קשת נבחרים בכל שלב על סמך הדרגות העדכניות שלהם.

השוני בתהליך הבנייה לעומת התהליך הבנייה המקורי נבע מהרצון לאפשר שליטה מדויקת בכמות הקשתות הסופית המתקבלת ברשת. דבר אשר איננו אפשרי לפי ההגדרה המקורית ( וגם לא במודל מס' 3 ), במידה ומעוניינים שהרשת תכיל מספר מסוים של קודקודים וקשתות.

הרשתות המתקבלות הינן בעלות תכונות דומות לרשתות המתקבלות באמצעות מודל מס' 3. השינוי שנעשה בתהליך הבניה, לא משפיע על תכונות אלו, הואיל והרשתות הנבנות באמצעותו הינן תת מחלקה של הרשתות המתקבלות ע"י המודל המקורי. (ניתן לראות זאת בצורה פשוטה ע"י האבחנה שהמודל המקורי יכול ליצור רשתות אלו, במידה וכל הקודקודים נוצרו בהתחלה ואח"כ נוצרו קשתות ).

### מודל העתקה דינמית

מודל זה מייצג את משפחת המודלים בעלי מנגנון ההעתקה. המודל הוצע ע"י Vasquez (Vazquez, 2002 ) ונועד לתאר את התפתחותן של רשתות ציטוטים. זהו מודל התפתחותי בו בכל שלב מתווסף קודקוד חדש ומקושר לקודקודים הקיימים. בניגוד למודלים המתבססים על "קישור מועדף" במודל זה לכל קודקוד המתווסף לרשת יש רק מידע חלקי על הקודקודים הקיימים, כאשר הוא מגלה מידע חדש על הרשת באמצעות השכנים שלו.

**הפרמטרים של המודל**

למודל זה יש פרמטר בודד שהינו ההסתברות לקישור הקודקוד החדש בקשת לאחד השכנים של הקודקוד אליו נוצר קשר.

**תהליך בניית הרשת באמצעות המודל מבוצע באופן הבא:**

**שלב א'**

יצירת רשת בעלת קודקוד בודד

**שלב ב'**

הוספת קודקוד לרשת וקישור שלו לאחד הקודקודים הקיימים ברשת - אשר נבחר באופן אקראי. מוסיפים את הקודקוד לרשימת הקודקודים שיש לעבור עליהם. עוברים על רשימת הקודקודים שבנינו ועבור כל קודקוד ברשימה בוחנים את רשימת הקשתות המקושרות אליו. עבור כל קשת מחליטים בהסתברות האם להעתיק את הקשת – כלומר האם ליצור קשת מהקודקוד החדש, לקודקוד אליו מקשרת קשת זו את הקודקוד מהרשימה . במידה ובחרנו להעתיק את הקשת, מוסיפים את קודקוד היעד של הקשת לרשימת הקודקודים שיש לעבור עליהם. לאחר שסיימנו לעבור על כל הקשתות של קודקוד מסוים מסירים את הקודקוד מהרשימה. תהליך זה נמשך עד אשר רשימת הקודקודים למעבר מתרוקנת.

חוזרים על שלב ב' עד ליצירת רשת בעלת כמות הקודקודים הנדרשת.

על מנת ליצור רשתות דומות לרשתות האמפיריות, מתבצעת חזרה על שלב ב' עד אשר כמות הקודקודים ברשת תהיה זהה לכמות הקודקודים ברשת האמפירית.

במסגרת עבודה זו ערך הפרמטר נבחר להיות 0.5.

### מודל "העולם הקטן" – גרסה מכוונת

זוהי הגרסה המכוונת של מודל העולם הקטן שהוצע ע"י Watts (Watts, 1999). מודל זה הוצע כהסבר אפשרי לתופעת "העולם הקטן" שהתגלתה ברשתות אמיתיות ומכאן נובע שם המודל. הסבר נוסף על המודל מופיע בסקירה.

**הפרמטרים של המודל:**

* + – כמות הקודקודים ברשת.
  + – דרגת כניסה והיציאה של כל קודקוד.
  + – פרמטר השולט על ההסתברות לחיווט מחדש של קשת.

**תהליך בניית הרשת באמצעות המודל מבוצע באופן הבא:**

**שלב א' – שלב יצירת הרשת.**

בונים סריג טבעתי בעל קודקודים. כך שכל קודקוד מחובר בקשת מכוונת ל-k הקודקודים הקרובים אליו ביותר. באופן מעשי, בצענו זאת ע"י מספור הקודקודים מ-0 עד , וקישור הקודקודים באופן הבא: עבור קודקוד שמספרו i, נוסיף קשתות מהקודקוד לכל קודקוד j המקיים: (במידה ו-k אי-זוגי מקשרים קודקוד נוסף מאחד הצדדים שנבחר באפן שרירותי).

**שלב ב' – שלב החיווט מחדש**

בשלב זה מתבצע מעבר על כל הקשתות ברשת. עבור כל קשת מחליטים בהסתברות p האם לשנות את קודקוד היעד שלה. במידה והוחלט לעשות כן, בוחרים באקראי קודקוד יעד חדש ומקשרים את הקשת אליו. עם זאת ישנה הגבלה על תהליך זה כך שלא ייווצרו קשתות עצמיות או קשתות כפולות. תהליך זה גורם לכך שברשת המתקבלת נוצרים "קיצורי דרך" הגורמים לכך שהמרחק הממוצע בין כל זוג קודקודים יהיה קטן, או במילים אחרות שהרשת הנוצרת תהיה בעלת תכונת "העולם הקטן" ( דבר זה תלוי גם בערכו של פרמטר החיווט מחדש p ). כמו-כן הרשת הנוצרת הינה בעלת ערך מקדם clustering גבוה עקב ההתבססות על הסריג ההתחלתי.

## מודלים לבחינת עוצמת הקשרים ברשתות

בחינת עוצמת הקשרים ברשתות בוצעה תוך שימוש במודלים של יצירת רשתות בינאריות אשר הותאמו ליצירת רשתות ממושקלות. דבר זה בוצע עקב הרצון לבחון האם התוצאות שהתקבלו עבור בחינת הקישוריות ניתנות להכללה כאשר עוברים לבחון גם את עוצמת הקשרים. לשם כך נבחנו ארבעה מודלים המייצגים מנגנונים שונים ליצירת רשתות אקראיות מכוונות, אשר מבוססים על המודלים שנבחנו בבחינת הקישוריות. עבור מודלים אלו בוצעו התאמות על מנת שייצרו רשתות ממושקלות. בנוסף, במהלך הבדיקה,נבחן מודל ביניים נוסף המשלב מנגנונים של שני מודלים קיימים. המודלים שנבדקו הם:

### מודל הגרף האקראי הממושקל המכוון

מודל זה הינו הגרסה הממושקלת של מודל הגרף האקראי המכוון אשר תואר בחלק הקודם. ההבדל היחיד בין גרסה זו לבין הגרסה המקורית הינו בכך שבגרסה זו ניתן ליצור יותר מקשת אחת בין זוג קודקודים. כתוצאה מן השינוי, בניית הרשת מבוצעת באופן שונה קמעה מאשר במקרה הרגיל. אם כי השיטה בה נוצרת הרשת במקרה זה ניתנת ליישום גם במקרה רגיל.

**בניית הרשת מבוצעת באופן הבא:**

* + קובעים את כמות הקודקודים N ואת סכום משקלי הקשתות – ברשת.
  + מייצרים רשת בעלת קודקודים ללא קשתות.
  + כל עוד סכום משקלי הקשתות ברשת שנוצרה קטן מ- מבצעים את הצעדים הבאים:
  + בוחרים זוג קודקודים שונים ברשת באופן אקראי ובלתי תלוי.
    1. במידה ולא קיימת קשת מכוונת בין זוג הקודקודים שנבחר ( קשת מכוונת מ- ל- ), מוסיפים לרשת קשת שכזו ונותנים לה את המשקל 1.
    2. אחרת מגדילים את המשקל של הקשת המכוונת מ- ל- הקיימת ב-1.

**אופן בחירת הפרמטרים עבור המודל**

על מנת ליצור רשתות הדומות לרשתות האמפיריות נבחר את להיות שווה למספר הקודקודים ברשת האמפירית ואילו ערכו של יהיה סכום משקלי הקשתות ברשת האמפירית.

### מודל "קישור מועדף" סטטי ממושקל

מודל זה הינו הרחבה של מודל ה"קישור מועדף" סטטי אשר תואר בחלק הקודם. הפרמטרים של מודל זה זהים לפרמטרים של המודל המקורי. השוני היחיד בין מודל זה למודל המקורי הינו בכך שכאשר המיקום של הקשת להוספה הינו בין זוג קודקודים שכבר מקושרים בקשת, אזי לא מתעלמים מהבחירה כמו במקרה הבינארי אלא מגדילים את משקל הקשת ב-1.

**אופן בחירת הפרמטרים עבור המודל**

על מנת ליצור באמצעות המודל רשתות דומות לרשתות האמפיריות, נבחר את להיות שווה למספר הקודקודים ברשת האמפירית ואילו הפרמטר ייבחר כך שערכו יהיה שווה לסכום משקלי הקשתות ברשת האמפירית.

בדומה למקרה הבינארי השתמשנו בערך 0.5 עבור הפרמטרים ו-.

### מודל "קישור מועדף" דינמי ממושקל

מודל זה הינו הרחבה של מודל ה"קישור מועדף" דינמי אשר תואר בחלק הקודם. הפרמטרים של מודל זה זהים לפרמטרים של המודל המקורי. השוני היחיד בין מודל זה למודל המקורי הינו בכך שכאשר נבחרים זוג קודקודים שכבר מקושרים בקשת, אזי מגדילים את משקל הקשת ב-1.

תהליך בניית הרשת נמשך עד אשר מספר הקודקודים וסכום משקלי הקשתות שהרשת מכילה מתאים לערכים של הרשת האמפירית.

*ערכי הפרמטרים בהם נעשה שימוש במקרה זה, נבחרו להיות דומים לערכים שנבחרו בבדיקת המקרה הבינארי.*

### מודל "קישור מועדף" דינמי ממושקל משופר

מודל זה הינו הרחבה של מודל ה"קישור מועדף" דינמי משופר אשר תואר בחלק הקודם. כמו במודלים הקודמים גם כאן השוני היחיד בין המודל המקורי לגרסה הממושקלת הינו בכך שכאשר נבחרים זוג קודקודים שכבר מקושרים בקשת, אזי מגדילים את משקל הקשת ב-1.

תהליך בניית הרשת מתמשך עד אשר מספר הקודקודים וסכום משקלי הקשתות שהרשת מכילה מתאים לערכים של הרשת האמפירית.

*ערכי הפרמטרים בהם נעשה שימוש במקרה זה, נבחרו להיות דומים לערכים שנבחרו בבדיקת המקרה הבינארי.*

### מודל "קישור מועדף" סטטי דינמי ממושקל

מודל זה משלב את המנגנונים של מודל ה"קישור מועדף" סטטי ממושקל ( מודל מס' 2 ) ומודל "קישור מועדף" דינמי ממושקל ( מודל מס' 3 ). מודל זה איננו חלק מקבוצת המודלים הראשונית שנבדקה. הסיבה לשימוש במודל זה הייתה בניסיון לשפר את איכות ההתאמה של המודל המנצח. פירוט על כך מופיע בהמשך.

* כפי שניתן לראות עבור שניים מהמודלים שנבחנו במקרה הבינארי - מודל העולם הקטן ומודל העתקה דינמית - לא ייצרנו גרסה ממושקלת. דבר זה נבע משני גורמים:

1. בניגוד לשאר המודלים, עבור מודלים אלו **אין** הרחבה טבעית למקרה הממושקל.
2. מודלים אלו הראו חוסר התאמה מובהק לנתוני הרשתות האמפיריות במקרה הבינארי. לכן הסבירות שהם יהוו את המודל המתאים ביותר במקרה הממושקל נמוך.

# מתודולוגיה

על מנת לזהות את המנגנון האחראי על יצירת הרשתות המקוונות נעשה שימוש במתודולוגיה הבאה:

## בניית מסווגים

עבור כל אחד מהמודלים ליצירת רשתות שנבחרו בוצעו הפעולות הבאות:

* כתיבת קוד המסוגל לחולל רשתות השייכות למודל בהתאם לפרמטרים הנדרשים.
* יצירת מספר רשתות לדוגמא בהתאם לפרמטרים של הרשתות האמפיריות.
* חישוב ווקטור מאפיינים רב-מימדי בשיטה שהוצעה ע"י Middendorf et al (2004 ), המתוארת בנספח א' , עבור כל אחת מהרשתות שנוצרו.

**עבור כל רשת אמפירית וכל זוג מודלים נבנה מסווג בינארי באופן הבא:**

* חלוקת ווקטורי המאפיינים של זוג המודלים לקבוצות אימון וקבוצות מבחן.
* בניית מסווג בינארי בעל רמת דיוק גבוהה עבור כל קבוצת מבחן,
* בדיקת איכות המסווג באמצעות קבוצת הבדיקה המתאימה, כאשר עבור כל מסווג חושבו הערכים המופיעים בטבלה 1.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **טבלה 1- תיאור חלוקת תוצאות בדיקת המסווג**   |  |  |  | | --- | --- | --- | | **מודל/ כמות** | **סיווג נכון** | **סיווג שגוי** | | מודל א | A | B | | מודל ב | C | D | |  |  |

* בוצעה חזרה על התהליך מספר פעמים עם קבוצות אימון ובדיקה שונות,על מנת לוודא כי איכות המסווג שנבנו, אינה נובעת כתוצאה מקבוצת האימון הספציפית.
* בוצע סיווג של הרשתות האמפיריות באמצעות כלל המסווגים המתאימים.

כפי שניתן לראות בטבלה 1 מתקיימות ההגדרות הבאות:

* הדיוק של המודל הינו אחוז הרשתות הכללי שסווגו נכונה מתוך קבוצת המבחן, ולכן שווה ל- .
* הסיכוי כי רשת שסווגה למודל **א**' סווגה **נכון** שווה ל-
* הסיכוי כי רשת שסווגה למודל **א**' סווגה **לא נכון** שווה ל-
* הסיכוי כי רשת שסווגה למודל **ב**' סווגה **נכון** שווה ל-
* הסיכוי כי רשת שסווגה למודל **ב**' סווגה **לא נכון** שווה ל-

## בדיקת עמידות המסווגים

מטרת המחקר הייתה לקבוע מי מבין המודלים שנבחרו מדמה באופן המדויק ביותר רשת אמפירית נתונה. עם זאת, ייתכן כי אף אחד מהמודלים שבחרנו מראש הינו המודל הנכון. במקרה זה הפרשנות של תוצאות הסיווג הינה מי מבין המודלים הכי פחות שגוי.

במידה והמודל שנבחר אכן מדמה באופן מדויק את הרשת האמפירית, אזי עבור כל תת קבוצה של מאפיינים של הרשת שנבחר, התוצאות שהתקבלו יראו כי מודל זה הינו המודל בעל הסבירות הגבוהה ביותר. אולם, ייתכן כי אף אחד מהמודלים אינו מדמה במדויק את הרשת האמפירית והמודל שנבחר הוא רק המודל הכי פחות שגוי. במקרה זה ייתכן כי חלק מהמאפיינים של הרשת האמפירית התאימו למודל אחד בעוד שמאפיינים אחרים התאימו למודל אחר.

לכן, אימנו מסווגים על תתי קבוצות שונות של מאפיינים ובדקנו האם בכל פעם התקבלה אותה התוצאה. במידה ודבר זה התרחש, אזי הנחנו, בסבירות גבוהה, כי סיווג הרשת האמפירית למודל אינו תלוי בקבוצת המאפיינים ואכן מתאר במדויק את הרשת האמפירית. על כן הגדרנו **מסווג עמיד** כמסווג אשר נותן תוצאות עקביות על תתי קבוצות של מאפיינים עליהם הוא מאומן. **ערך העמידות** של המסווג הוגדר כאחוז התוצאות העקביות מתוך כלל הסיווגים שבוצעו על תתי קבוצות מאפיינים.

## בחירת המסווגים

על מנת שהסבירות של תוצאות הסיווג תהיה גבוהה השתמשנו רק במסווגים אשר עמדו בתנאים הבאים:

1. ערך העמידות של המסווג היה גבוה.
2. תוצאת סיווג הרשת האמפירית ע"י המסווג הכולל – המסווג על כלל המאפיינים הייתה זהה לתוצאת הסיווג של המסווגים המתאימים על תתי קבוצות המאפיינים.

## מציאת המודל הזוכה

על מנת לגלות מהו המודל המדמה באופן הטוב את הרשתות, פעלנו במספר שלבים:

1. בדקנו עבור כל רשת אמפירית, מיהו המודל אשר "ניצח" את כל שאר המודלים, כלומר מיהו המודל אשר אליו סווגה הרשת האמפירית **בכל** אחד מהמסווגים הבינאריים בהם הוא השתתף.
2. לאחר מציאת המודל המנצח, בדקנו האם קיים מודל אשר ניצח את שאר המודלים **ברוב מובהק** של הרשתות שנבדקו.
3. עבור כל מודל שכזה, בדקנו את המובהקות הסטטיסטית של ההשערה כי מודל זה אכן מדמה את הרשתות האמפיריות טוב יותר מאשר שאר המודלים, כלומר בדקנו כי ההסתברות שהתוצאות שקיבלנו התקבלו באקראי נמוך מערך סף שקבענו. (השאיפה, כמובן, הייתה שערך הסף יהיה נמוך ככל האפשר ).

## בדיקת איכות תוצאות הסיווג

המסווגים שנבנו עשויים לעיתים לשגות ולסווג את הרשתות האמפיריות למודל הלא נכון. רצינו לקבל הערכה של איכות הסיווג, כפונקציה של אחוז הרשתות האמפיריות שסווגו נכונה. על מנת לבצע זאת התבססנו על שני גורמים:

1. הערכת הדיוק ( ה-Precision ) של המסווגים שחושבה בעת בניית המסווגים כפי שמתואר בסעיף ‎6.1.
2. שימוש בעובדה שהסיווג של הרשתות האמפיריות בין זוגות המודלים בוצע עבור מספר רב של רשתות אמפיריות.

**בדיקת האיכות של תוצאות הסיווג בוצעה באופן הבא:** עבור כל סוג מסווג חושבה הסתברות המאורע שכלל תוצאות הסיווג שלו על כל הרשתות עליהם הוא הופעל נבעו כתוצאה מסיווג נכון שלו. הואיל והתוצאות הן בלתי תלויות, אזי ההסתברות שווה למכפלת ההסתברויות שכל תוצאה בפני עצמה נבעה כתוצאה מסיווג נכון של המסווג. לכן קיבלנו כי ההסתברות שווה ל: , כאשר הינה קבוצת כלל תוצאות הסיווגים של הרשתות האמפיריות ו- הינה ההסתברות כי הסיווג של הרשת האמפירית ה- היה נכון. ערכו של תלוי בתוצאת הסיווג של התוצאה ה-. כל מסווג הוא בינארי – מבחין בין שני מודלים: א' ו-ב', כך שאם הרשת האמפירית ה- סווגה למודל א' אזי נקבל כי , אחרת נקבל כי .

בדיקה זו בלבד לא הספיקה על מנת להראות כי תוצאות הסיווג של מסווג ספציפי הינן באיכות טובה. דבר זה נובע משני גורמים: מחד, בחלק נכבד מהמקרים, איכות המסווגים , לא הייתה מושלמת, ומאידך הבדיקה בוצעה על כמות גדולה של תוצאות. לכן עקב היות איכות המסווגים מספר חיובי הקטן מאחד, תוצאת המכפלה שלהם הינה נמוכה משמעותית מהסף שקבענו, זאת למרות שכל מסווג בפני עצמו היה בעל איכות סיווג גבוהה. למשל במקרה שאיכות כל אחד מהמסווגים הינה 95%, וקיימים 20 תוצאות סיווג, אזי תוצאת החישוב נותנת ערך של  *, ערך הנמצא במרחק רב מאוד מהסף אותו קבענו.*

*לכן, חישבנו גם את ההסתברויות בהם חלק מהמסווגים טעו, אך עדיין המודל המנצח זוכה באחוז משמעותי של המקרים ולפחות ברובם. למשל אם בדקנו 20 רשתות, וקיבלנו כי המסווגים הכריעו לטובת מודל א' ב-15 מהמקרים, ובשאר המקרים הכריעו לטובת מודל ב', נחשב את סכום ההסתברויות לכך שהיו לכל היותר ארבע טעויות בסיווג לטובת מודל א', כך שגם במקרה זה מודל א' זוכה לפחות ב-11/20 מהמקרים. סיכמנו את ההסתברויות שחישבנו ובדקנו אם הסכום שהתקבל שנסמנו ב- גבוה מערך הסף שנקבע. במקרה זה אמרנו כי באופן מובהק מודל א' טוב יותר מאשר מודל ב' כמודל המייצג את הרשתות האמפיריות עם הסתברות שגיאה השווה* ל- *.*

מובן כי שאיפתנו המקורית הייתה להגיע למצב בו ציון המובהקות הנדרש יתקבל עבור זכייה של המודל הזוכה לא ברוב רגיל של יותר מ-50% מהמקרים, אלא ברוב מכריע גדול ככל הניתן.

# תוצאות

המתודולוגיה שתוארה בחלק הקודם הופעלה על נתוני הרשתות האמפיריות. יישום המתודולוגיה בוצע באמצעות תוכנה שפותחה למטרת מחקר זה ומתוארת בנספח ב', וכן באמצעות שימוש בכלים חיצוניים הניתנים לשימוש חופשי ברשת האינטרנט אשר יצוינו בהמשך.

## תוצאות בחינת קישוריות הרשתות

**על מנת לאפשר הצגה נוחה יותר של התוצאות, בהמשך העבודה ייעשה שימוש בקיצורים הבאים עבור המודלים השונים.**

1. מודל הגרף האקראי המכוון – DRG.
2. מודל "קישור מועדף" סטטי – SPA.
3. מודל "קישור מועדף" דינמי – DPA.
4. מודל "קישור מועדף" דינמי משופר – IDPA.
5. מודל העתקה דינמית – DC.
6. גרסה מכוונת של מודל "העולם הקטן" – DSW.

### ביצוע הסיווג

על מנת לסווג את הרשתות האמפיריות, חוללו דוגמאות אימון מהמודלים שתוארו לעיל בעלות מספר קודקודים זהה למספר הקודקודים ברשתות האמפיריות. הואיל וכמות הקשתות בחלק מהמודלים לא ניתנת לקביעה מדויקת, נבחרו מתוך הדוגמאות שחוללו רק רשתות אשר מספר הקשתות שלהן סטה לכל היותר ב-2% מממספר הקשתות ברשתות האמפיריות.

עבור כל רשת אמפירית, חולל אנסמבל של 500 רשתות-דוגמא, עבור כל מודל, לאימון המסווגים. לאחר יצירת רשתות-הדוגמא, חולל עבור כל אחת מהן ווקטור תכונות רב-מימדי בעל 4680 מימדים ע"פ השיטה שהוצעה ע"י Middendorf et al המתוארת בנספח א'. כמות זו של מאפיינים מבוססת על תוצאות הבחינה במאמר של Middendorf et al, כי זוהי כמות מספקת ליצירת מסווגים בעלי דיוק גבוה, כפי שאכן היה בפועל גם במקרה שלנו. בנוסף בוצעה בדיקה עצמאית שהראתה כי חילול של פחות מאפיינים יצר מסווגים בעלי רמת דיוק נמוכה יותר.

* למרות זאת, כפי שניתן לראות מתוצאות בדיקת העמידות בהמשך, אין בפועל צורך בכל המאפיינים שנוצרו, אולם לא ניתן לייצר מראש אך ורק את המאפיינים הרלבנטיים באמצעות שימוש בשיטת Middendorf, ולכן נדרשה כמות זו של מאפיינים.

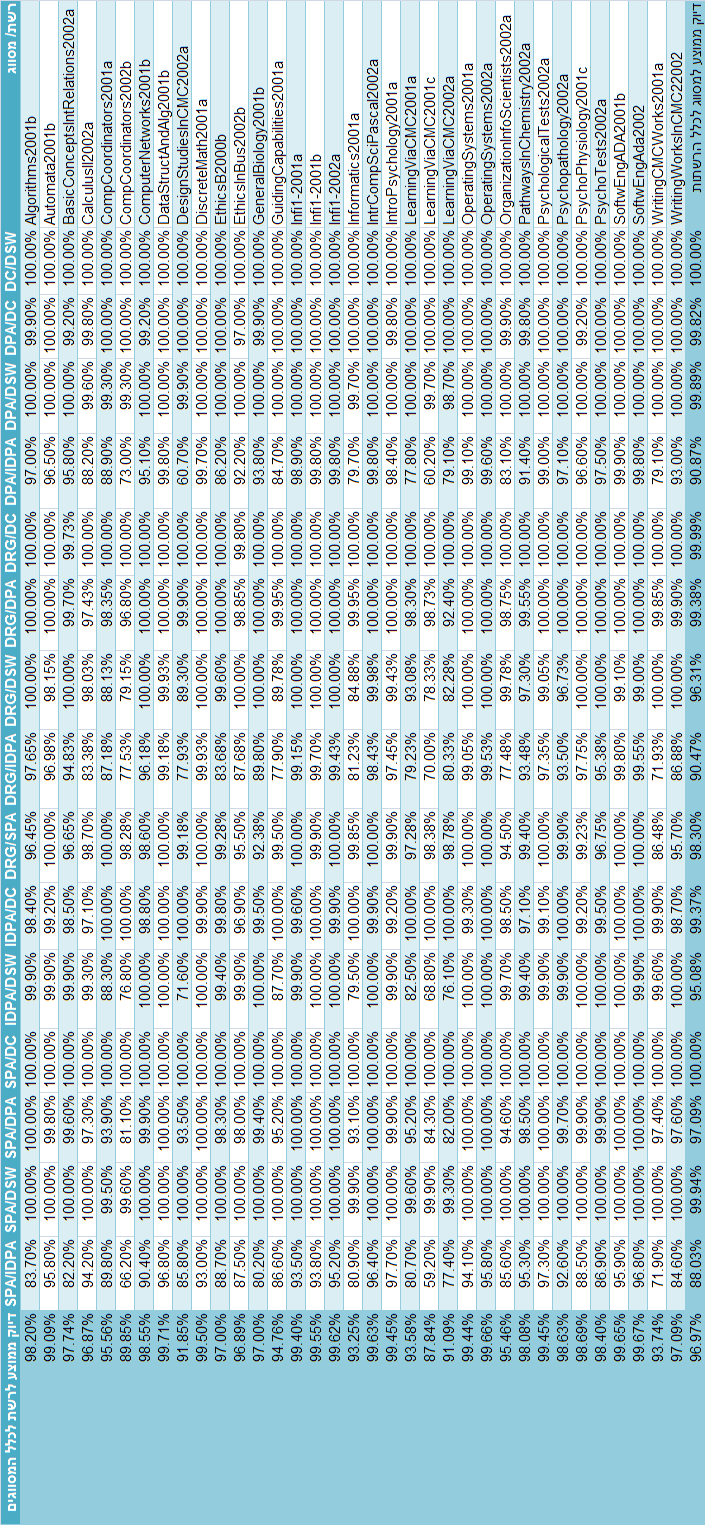
אנסמבל רשתות-הדוגמא חולק לשתי קבוצות: קבוצת אימון וקבוצת מבחן. החלוקה בין קבוצות האימון והמבחן הייתה ביחס של 1:4, כלומר 1/5 מרשתות הדוגמא שימשו לאימון המודל והשאר לבדיקת הדיוק שלו.

עבור כל זוג מודלים, אומן מסווג בעל דיוק גבוה מסוג Support Vector Machine (Vapnik, 1995). בניית המסווג בוצעה, על סמך נתוני קבוצת האימון, באמצעות מימוש חופשי (freeware) בשפת C בשם SVM-Light (Joachims, 1999). ביצועי המסווג נבדקו על קבוצת המבחן על מנת לוודא כי המסווג שנבנה בעל איכות טובה, ולחשב את הדיוק שלו. תהליך זה בוצע 5 פעמים על קבוצות אימון ומבחן שונות ( 5-fold Cross Validation ), על מנת לוודא כי אין תנודות קיצוניות באיכות המסווגים שאומנו על קבוצות האימון השונות. בסיומו של התהליך נבדקו הן התוצאות המשוקללות של כלל המסווגים שנוצרו והן תוצאות הסיווג של המסווג בעל אחוזי הדיוק הגבוהים ביותר.

* יש לציין כי ביישומים דומים של Machine Learning היחס בין גדלי קבוצות האימון והבדיקה בד"כ שונה וקבוצת האימון גדולה מקבוצת המבחן. דוגמא לכך ניתן לראות ב-Middendorf et al (2004) בו בוצעה בחינה דומה במהלכה בוצע שימוש ב-80% מהדוגמאות לצורך האימון. בעבודה זו, נבחרה מלכתחילה קבוצת אימון קטנה יחסית של כ-20% מהדוגמאות. דבר זה בוצע מתוך מטרה להקטין את החשש של התאמת יתר ( Over fitting ) של המסווגים. התאמת יתר הינה בעיה מוכרת בתחום של Machine-Learning, אשר עשויה, במקרה של בניית מסווגים, להביא ליצירת מסווגים המותאמים יתר על המידה לדוגמאות הספציפיות שניתנו במסגרת האימון, אך אינם בעלי יכולת הכללה טובה.
* יתרון נוסף לשימוש בקבוצת אימון קטנה, הינו בכך שנותרת קבוצת מבחן גדולה יותר, אשר מאפשרת קבלת הערכה טובה יותר של איכות המסווגים, הבאה לידי ביטוי באחוז הסיווגים הנכונים בקבוצת המבחן. ובכך מאפשרת מתן ציון מדויק יותר עבור התוצאות.

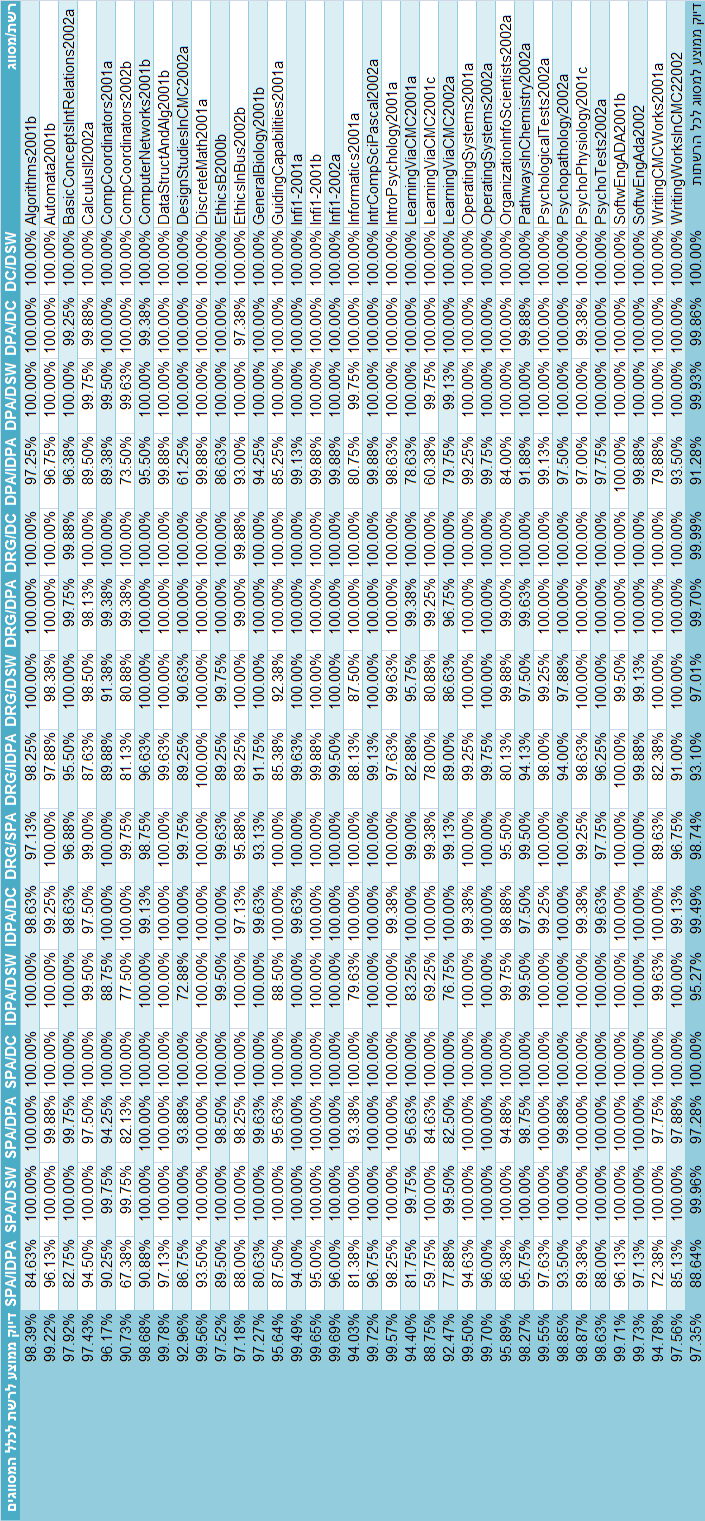
באופן מעשי מתברר כי למרות השימוש בקבוצת האימון הקטנה, המסווגים שנוצרו היו בעלי איכות גבוהה, כפי שניתן לראות בטבלאות 1א ו-1ב.

**טבלה 1א – אחוזי דיוק ממוצעים של המסווגים עבור המקרה הבינארי**



**טבלה 1א:** טבלה זו מכילה את אחוזי הדיוק הממוצעים (הממוצע של חמשת המסווגים שנבנו מכל סוג לכל רשת ) של המסווגים השונים עבור כל אחת מהרשתות כפי שהתקבלו עבור המקרה הבינארי.

***טבלה 1ב – אחוזי דיוק של המסווגים הטובים ביותר עבור המקרה הבינארי***

****

**טבלה 1ב:** טבלה זו מכילה את אחוזי הדיוק של המסווגים בעלי אחוזי הדיוק הטובים ביותר עבור כל אחת מהרשתות.

כפי שניתן לראות בטבלאות 1א' ו-1ב' אחוזי הדיוק של המסווגים השונים גבוהים מאוד, ועומדים ברוב המקרים על מעל 98%. עם זאת ישנה חריגה הבולטת לעין במקרה של המסווגים שמטרתם לסווג בין המודלים השונים לבין המודל IDPA. עבור מסווגים אלו אחוזי הדיוק נמוכים יותר ובמקרה הגרוע ביותר עומדים **רק** על כ-88% בממוצע. הסיבה לכך איננה ברורה לחלוטין. עם זאת נראה כי המסווגים בעלי אחוזי הדיוק הנמוכים יחסית התרחשו במספר מצומצם של רשתות שנבדקו, ולכן ייתכן כי הקושי בסיווג רשתות אלו, נובע לא רק מהקושי של המסווג להבדיל בין המודלים השונים אלא גם כתוצאה מהפרמטרים של הרשת ( כמות הקודקודים והקשרים ).

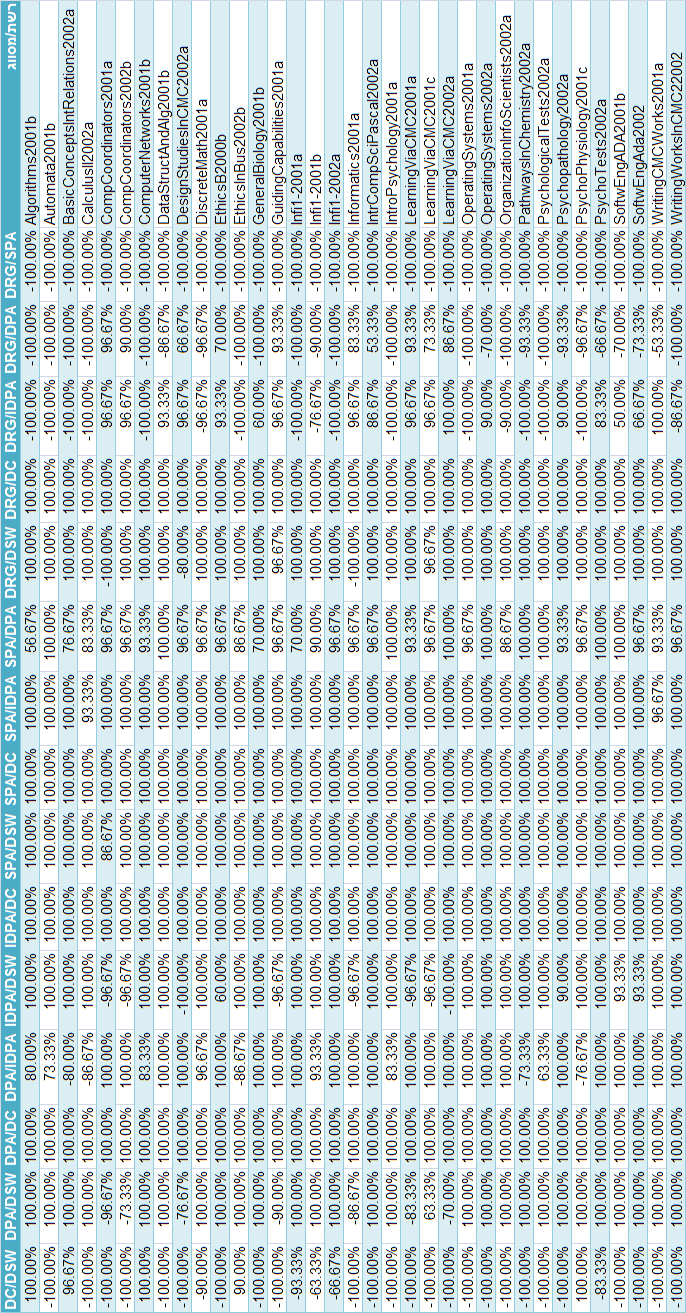
### בדיקת עמידות המסווגים

עבור כל זוג מודלים בוצעה בדיקה של עמידות המסווגים. הבדיקה בוצעה באמצעות בחירה של 30 תתי קבוצות בעלות 500 מאפיינים כל אחת ( כ-10% מכמות המאפיינים הכוללת ) אשר נבחרו באופן אקראי. עבור כל תת קבוצה נבנה מסווג מסוג SVM, ונבדקו תוצאות הסיווג שלו על הרשת האמפירית המתאימה. בתרשים 2 מוצגת היסטוגרמה המתארת את ערכי העמידות של חמישה מן המסווגים, ובטבלה 2 מוצגים הערכים עליהם מתבססת ההיסטוגרמה ביתר פירוט. (טבלה 2 מכילה את ערכי העמידות של כל המסווגים, ולא רק של החמישה המוצגים בהיסטוגרמה ).

**תרשים 2 – היסטוגרמת עמידות המסווגים עבור המקרה הבינארי**

**תרשים 2:** התרשים מכיל היסטוגרמה המציגה את הכמויות השונות של ערכי העמידות של חמישה מן המסווגים שנבדקו. הציר האופקי הינו ערך העמידות והציר האנכי הינו כמות המסווגים מכל סוג ( מתוך 35 מסווגים ) שערך העמידות שלה היה בטווח.

**טבלה 2 – עמידות המסווגים עבור המקרה הבינארי**



**טבלה 2:** טבלה זו מכילה את פירוט תוצאות בדיקת העמידות של המסווגים השונים. הערך הרשום בתאים השונים הינו אחוז המסווגים אשר היו בעלי תוצאת סיווג זהה. ערך חיובי/שלילי בתא משמעותו כי הרשת סווגה אל המודל השמאלי/ימני בהתאמה כפי שמופיע בכותרת העמודה המתאימה.

כפי שניתן לראות, העמידות שגילו המסווגים הייתה גבוהה מאוד. עבור המסווגים שהוצגו בהיסטוגרמה למעלה מ-80% היו בעלי עמידות מושלמת, ועבור למעלה מ-95% הייתה עמידות של לפחות 90%. נתונים אלו דומים גם עבור כלל המסווגים: כמעט ב-80% מהמקרים המסווגים היו בעלי עמידות מושלמת ועבור למעלה מ-90% מהמקרים עמידות המסווג הייתה לכל הפחות 90%.

### זיהוי המודל המתאים ביותר

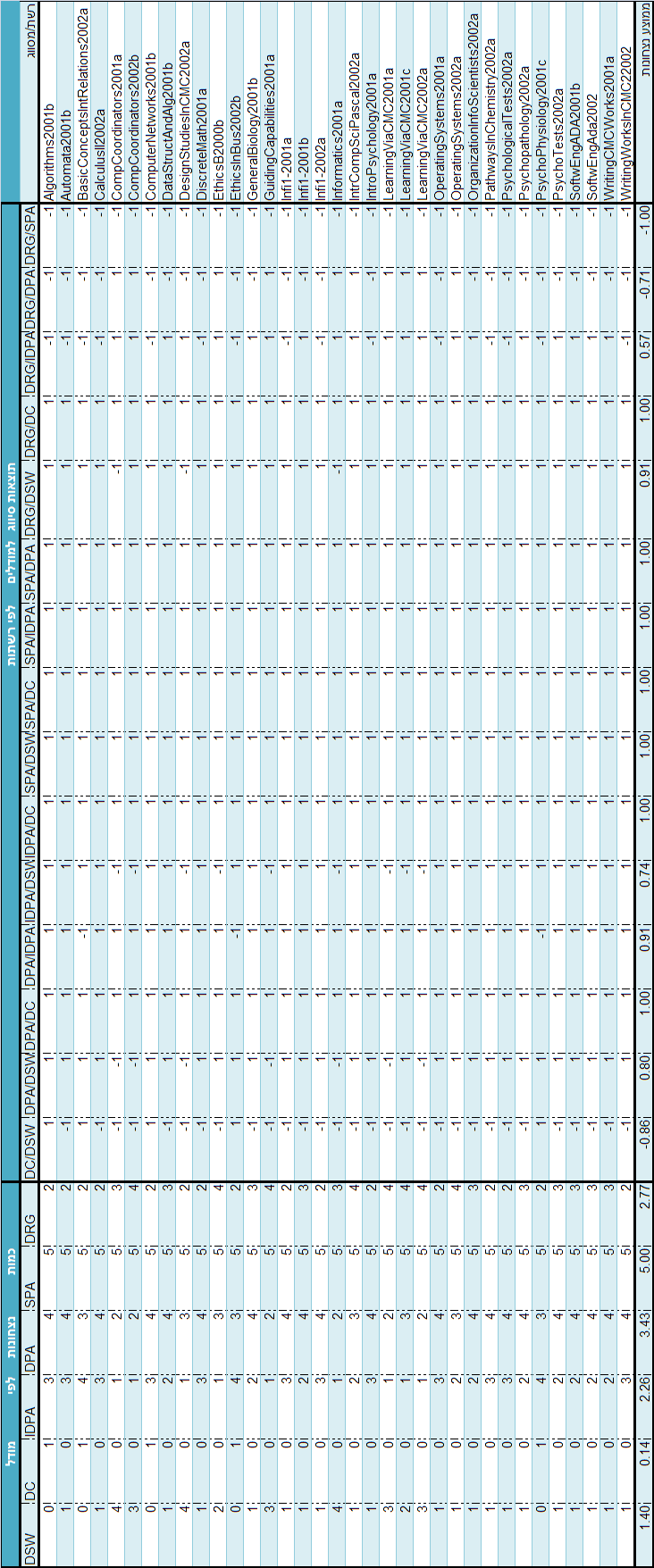
זיהוי המודל המנצח נעשה בשיטת ההצבעה. עבור כל מודל, נבדקה כמות המסווגים בהם סווגו הרשתות האמפיריות למודל. טבלאות 3א ו-3ב מכילות את ריכוז התוצאות ואת תוצאות הסיווג המפורטות בהתאמה. כפי שניתן לראות המודל "קישור מועדף" סטטי - SPA, היה בעל ההתאמה הטובה ביותר, באופן חד משמעי, כאשר הרחק מאחוריו משתרך המודל "קישור מועדף" דינמי - DPA.

**טבלה 3א – ריכוז תוצאות סיווג הרשתות האמפיריות למודלים.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **מסווג/מסווג** | **הצבעות** | **DRG** | **SPA** | **DPA** | **IDPA** | **DC** | **DSW** |
| DRG | 3/5 |  | -100.00% | -71.43% | 57.14% | 100.00% | 91.43% |
| **SPA** | **5/5** | **100.00%** |  | **100.00%** | **100.00%** | **100.00%** | **100.00%** |
| DPA | 4/5 | 71.43% | -100.00% |  | 91.43% | 100.00% | 80.00% |
| IDPA | 2/5 | -57.14% | -100.00% | -91.43% |  | 100.00% | 74.29% |
| DC | 0/5 | -100.00% | -100.00% | -100.00% | -100.00% |  | -85.71% |
| DSW | 1/5 | -91.43% | -100.00% | -80.00% | -74.29% | 85.71% |  |

**טבלה 3א:** טבלה זו מכילה בעמודה **הצבעות** את **מספר** המסווגים שסיווגו את הרשת למודל המופיע באותה שורה. בסה"כ נבדקו שישה מודלים, וכל מודל השתתף בחמישה מסווגים, ולכן מספר הניצחונות המקסימאלי הוא חמישה. שאר הערכים בטבלה מציגים את **אחוז** ההצבעות שקיבל המודל לעומת המודלים האחרים. במידה והערך חיובי הדבר מצביע על כך כי היו יותר מסווגים אשר סיווגו את הרשתות האמפיריות למודל המופיע בשורה המתאימה מאשר למודל המופיע בעמודה המתאימה ולהיפך. הערכים הם בטווח -100%-100%. לכן, למשל, אם הערך המופיע הוא 60% המשמעות הינה ש-80% מהמסווגים סיווגו את הרשתות האמפיריות למודל המופיע בשורה ולא למודל המופיע בעמודה. **כפי שניתן לראות המודל SPA ניצח את שאר המודלים בכל המקרים.**

**טבלה 3ב – תוצאות הסיווג המפורטות של רשתות למודלים**



**טבלה 3ב:** טבלה זו מכילה עבור כל מסווג את תוצאות הסיווג של הרשתות האמפיריות. ערך חיובי/שלילי בתא משמעותו כי הרשת סווגה אל המודל השמאלי/ימני בהתאמה כפי שמופיע בכותרת העמודה המתאימה. בסיום הטבלה מופיע הממוצע של כמות הפעמים בהם סווגו הרשתות האמפיריות למודל מסוים בכלל המסווגים בהם הוא השתתף. הואיל וכל מודל השתתף בחמישה מסווגים הרי שכמות הניצחונות המקסימאלית הינה 5.

### בדיקת איכות תוצאות הסיווג

על מנת לוודא כי התוצאות שהתקבלו אכן בעלות מובהקות סטטיסטית, כלומר הן לא התקבלו במקרה עקב שגיאות בסיווג, בוצעה בדיקה של מובהקות התוצאות כפי שתואר בחלק המתודולוגיה. במסגרת בדיקה זו, חושב עבור כל סוג מסווג, הקשר בין כמות הסיווגים השגויים של הרשתות השונות לבין ההסתברות של הסיווג. תוצאות הבדיקה עבור המסווגים בהם השתתף המודל המנצח ( SPA ). מוצגות בתרשים מס' 3, ואילו בטבלה 4 מוצגים הנתונים עליהם מתבסס התרשים.

ניתן לראות כי ברוב המקרים מתקבלות תוצאות בעלות הסתברות גבוהה, עבור מספר שגיאות לא גדול. למעשה, מספיק להניח כי למסווג היו לכל היותר 3 שגיאות בסיווג על מנת לקבל כי תוצאות הסיווג הינן בעלי הסתברות של למעלה מ-99.9%. הנחה של כמות שגיאות זו שקולה לכך שכ-90% מהרשתות סווגו נכונה (במקרה של המסווג SPA/DPA ) ועבור שאר המסווגים התוצאות אף טובות יותר.

המקרה היחיד בו התוצאות פחות טובות הינו עבור המסווג SPA/IDPA. כפי שצוין בסעיף 7.1.1 הדיוק של מסווג זה היה נמוך משמעותית לעומת הדיוק של שאר המסווגים ( 88% דיוק לעומת למעלה מ-95% דיוק ). דבר זה הביא להקטנת ההסתברות כי הסיווג של **כלל** הרשתות האמפיריות באמצעות מסווג זה נכונה. או לחילופין מחייב להניח שמספר שגיאות הסיווג היה רב יותר ( 10 שגיאות לעומת 3 שגיאות ), על מנת לקבל הסתברות סיווג של למעלה מ-95%. עם כל זאת, עדיין מדובר על כך שבהסתברות של למעלה מ-95% למעלה מ-71% מהרשתות ( 25 מתוך 35 רשתות ) סווגו נכונה למודל ה-SPA. למרות שתוצאות אלו נמוכות יחסית לשאר המסווגים עדיין יש בכך להעיד כי המודל SPA מייצג טוב יותר את הרשתות מאשר מודל ה-IDPA.

**תרשים 3- הקשר בין כמות טעויות הסיווג המירבית לבין הסתברות תוצאות הסיווג עבור המסווגים השונים.**

**תרשים 3:** בתרשים זה ניתן לראות כיצד ככל שמגדילים את כמות טעויות סיווג האפשרית, כך ההסתברות של הסיווג הכולל הולכת וגדלה.

**טבלה 4- נתונים מפורטים של הסתברות תוצאות הסיווג בהתאם לכמות השגיאות המירבית.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **כמות שגיאות/ מסווג** | **DRG/SPA** | **SPA/DPA** | **SPA/DC** | **SPA/DSW** | **SPA/IDPA** |
| 0 | 0.613 | 0.219 | 1.000 | 1.000 | 0.001 |
| 1 | 0.920 | 0.590 | 1.000 | 1.000 | 0.010 |
| 2 | 0.990 | 0.858 | 1.000 | 1.000 | 0.043 |
| 3 | 0.999 | 0.967 | 1.000 | 1.000 | 0.123 |
| 4 | 1.000 | 0.995 | 1.000 | 1.000 | 0.259 |
| 5 | 1.000 | 0.999 | 1.000 | 1.000 | 0.437 |
| 6 | 1.000 | 1.000 | 1.000 | 1.000 | 0.620 |
| 7 |  |  |  |  | 0.775 |
| 8 |  |  |  |  | 0.883 |
| 9 |  |  |  |  | 0.947 |
| 10 |  |  |  |  | 0.979 |
| 11 |  |  |  |  | 0.993 |
| 12 |  |  |  |  | 0.998 |

**טבלה 4:** טבלה זו מציגה את הסתברות הסיווג המדויקת עבור כל מסווג בהתאם לכמות השגיאות המירבית. ניתן לראות כי ברוב המקרים הגענו להסתברות גבוהה מאוד של 0.999 תוך מספר שגיאות מצומצם ביותר.

חשוב להזכיר כי לצורך בדיקת המובהקות בוצע שימוש רק במסווגים שערך העמידות שלהם עבר את הסף של 90%. הגבלה זו גרמה לכך כי בחלק מהמקרים הבדיקה בוצעה על מספר נמוך יותר של תוצאות. טבלה מס' 5 מכילה פירוט של כמות המסווגים שהיו בשימוש בבדיקת המובהקות.

**טבלה מס' 5- כמות המסווגים שנבדקו במסגרת בדיקת המובהקות**

|  |  |
| --- | --- |
| **מסווג** | **כמות תוצאות** |
| DRG/SPA | 35 |
| SPA/DPA | 28 |
| SPA/DC | 35 |
| SPA/DSW | 34 |
| SPA/IDPA | 35 |

**טבלה 5:** בטבלה ניתן לראות כי עבור רוב המסווגים, עמידות המסווגים הייתה מושלמת ובהתאם לכך היה שימוש בכמעט כל תוצאות. יוצא הדופן היחיד הינו המסווג SPA/DPA בו רק 80% ( 28 מתוך 35 ) מהמסווגים היו בעלי עמידות כנדרש.

## בחינת עוצמת הקשרים – Weighted Case

**על מנת לאפשר הצגה נוחה יותר של התוצאות, בהמשך העבודה ייעשה שימוש בסימונים הבאים עבור המודלים השונים:**

1. מודל הגרף האקראי הממושקל המכוון – WDRG.
2. מודל "קישור מועדף" סטטי ממושקל – WSPA.
3. מודל "קישור מועדף" דינמי ממושקל – WDPA.
4. מודל "קישור מועדף" דינמי משופר ממושקל – WIDPA.
5. מודל "קישור מועדף" סטטי דינמי ממושקל - WSDPA.

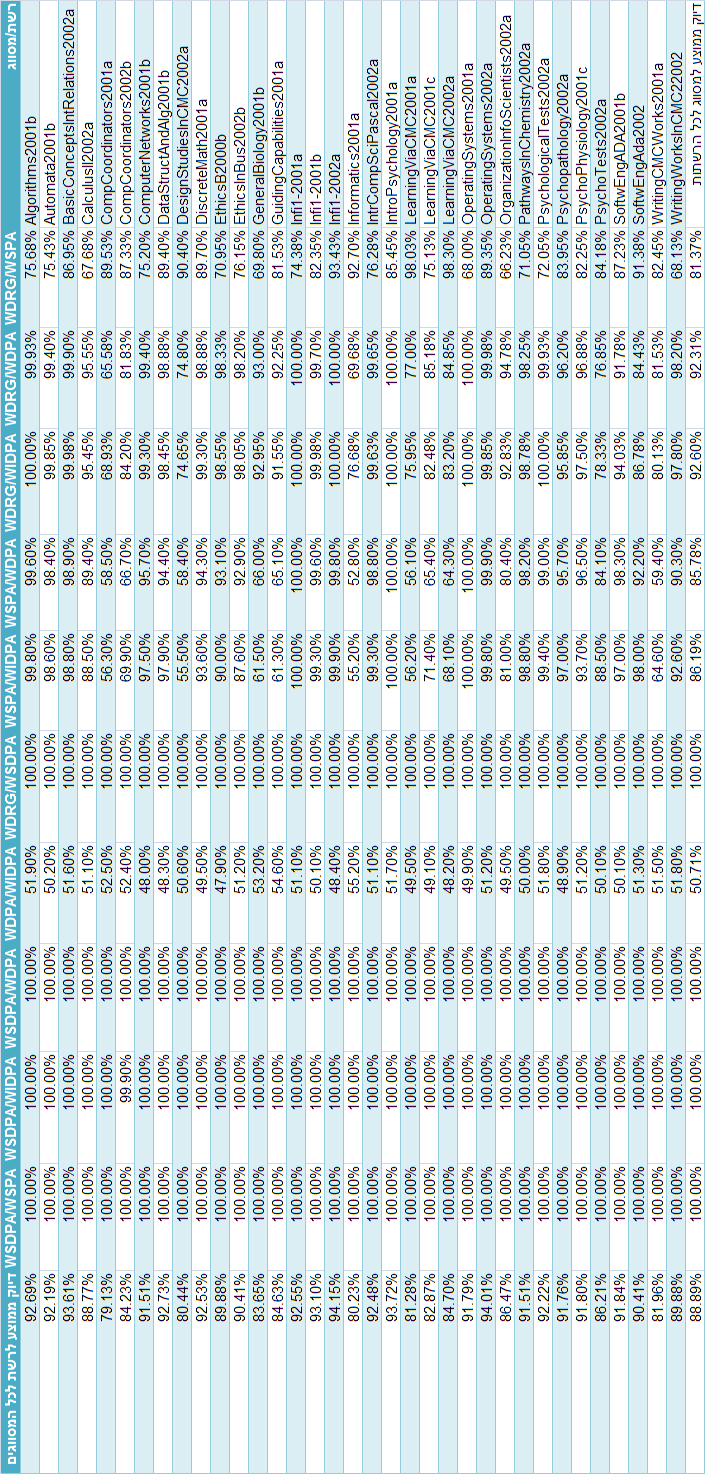
### ביצוע הסיווג

סיווג הרשתות האמפיריות במקרה הממושקל בוצע באופן דומה לבדיקת המקרה הבינארי. חוללו דוגמאות אימון מהמודלים שתוארו לעיל בעלי מספר קודקודים זהה למספר הקודקודים ברשתות האמפיריות. מתוך הדוגמאות שחוללו נבחרו רק רשתות אשר מספר הקשתות שלהן סטה לכל היותר ב-2% מממספר הקשתות ברשתות האמפיריות. עבור כל רשת אמפירית, חוללו 500 דוגמאות, עבור כל מודל, לאימון המסווגים. לאחר יצירת הרשתות, חולל עבור כל אחת מהן ווקטור תכונות רב-מימדי בעל 4680 מימדים ע"פ השיטה שהוצעה ע"י Middendorf et al המתוארת בנספח א'.

* יש לציין כי במקרה זה משמעות המאפיינים המרכיבים את הווקטור, שונה מאשר במקרה הבינארי. עם זאת במידה ומתייחסים לעוצמת הקשר בין זוג קודקודים כאל קיומם של מספר קשתות שונות בין הקודקודים, ובפועל זוהי אכן המשמעות של עוצמת הקשר ברשתות שנוצרו ע"ס המודלים, אזי המשמעות של המאפיינים זהה למשמעותן במקרה הבינארי.

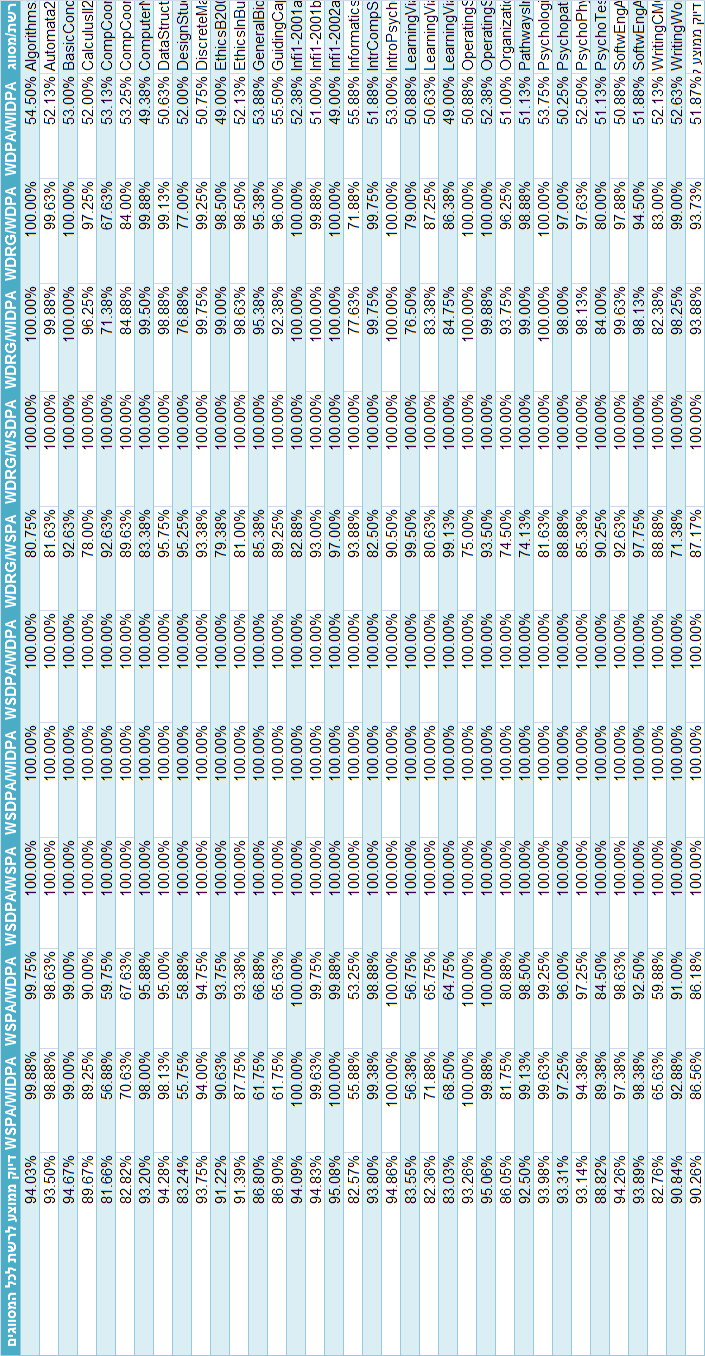
הדוגמאות חולקו לשתי קבוצות: קבוצת אימון וקבוצת מבחן. חלוקת הדוגמאות בין קבוצות האימון והמבחן הייתה ביחס של 1:4, כלומר 1/5 מהדוגמאות שימשו לאימון המודל והשאר לבדיקת הדיוק שלו. עבור כל זוג מודלים, אומן מסווג מסוג SVM באמצעות מימוש חופשי (freeware) בשם SVMLight. המסווג אומן על קבוצת האימון, וביצועיו נבדקו על קבוצת המבחן. על מנת לוודא כי המסווג שנבנה בעל איכות טובה, ולחשב את הדיוק שלו. לתהליך זה בוצע 5 פעמים על קבוצות אימון ובדיקה שונות ( 5-fold Cross Validation ), על מנת לוודא כי אין תנודות קיצוניות באיכות המסווגים שאומנו על קבוצות האימון השונות. בסיומו של התהליך נבדקו הן התוצאות המשוקללות של כלל המסווגים שנוצרו והן תוצאות הסיווג של המסווג בעל אחוזי הדיוק הגבוהים ביותר. בטבלאות 6א' ו-6ב' מציגות את רמת הדיוק של המסווגים שנבנו.

**טבלה 6א - אחוזי דיוק ממוצעים של המסווגים עבור המקרה הממושקל**

****

**טבלה 6א:** טבלה זו מכילה את אחוזי הדיוק הממוצעים (הממוצע של חמשת המסווגים שנבנו מכל סוג לכל רשת ) של המסווגים השונים עבור כל אחת מהרשתות כפי שהתקבלו עבור המקרה הממושקל.

**טבלה 6ב – אחוזי דיוק של המסווגים הטובים ביותר עבור המקרה הממושקל**



**טבלה 6ב:** טבלה זו מכילה את אחוזי הדיוק של המסווגים בעלי אחוזי הדיוק הטובים ביותר עבור כל אחת מהרשתות. עברו המקרה הממושקל.

כפי שניתן לראות בטבלאות, איכות המסווגים עבור מקרה זה הייתה נמוכה יותר מאשר בבחינת המקרה הבינארי. הדבר נובע בעיקר מכך שאיכות המסווג WDPA/IDPA הייתה גרועה מאוד ועמדה על כ-50%. למעשה דבר זה אומר שבניית המסווג WDPA/IDPA נכשלה. אולם גם בחלק משאר המסווגים האיכות שהתקבלה הייתה פחות טובה ועמדה על כ-86%-87%. עם זאת, למרות הירידה זו באיכות המסווגים, עדיין הצלחנו לקבל תוצאות בעלות מובהקות גבוהה כפי שיתואר בהמשך. ולכן ההשלכות של הירידה באיכות המסווגים על הבדיקה שנעשתה היו בעלות משמעות נמוכה.

### זיהוי המודל בעל ההתאמה הטובה ביותר

בדומה למקרה הבינארי זיהוי המודל המתאים ביותר נעשה בשיטת ההצבעה. עבור כל מודל, נבדקה כמות המסווגים אשר סיווגו את הרשתות האמפיריות למודל. הפעם בניגוד למקרה הבינארי לא היו תוצאות חד משמעיות. טבלה מספר 7 מכילה את ריכוז התוצאות.

**טבלה 7 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **מודל/מודל** | **הצבעות** | **WDRG** | **WSPA** | **WDPA** | **WIDPA** |
| WDRG | 0/3 |  | -100.00% | -80.00% | -77.14% |
| **WSPA** | **3/3** | **100.00%** |  | **62.86%** | **57.14%** |
| WDPA | 1/3 | 80.00% | -62.86% |  | -65.71% |
| WIDPA | 2/3 | 77.14% | -57.14% | 65.71% |  |

**טבלה 7:** טבלה זו מכילה בעמודה **הצבעות** את **מספר** המסווגים שסיווגו את הרשת למודל המופיע באותה שורה. בסה"כ נבדקו ארבעה מודלים. כל מודל סווג באמצעות שלושה מסווגים,ולכן מספר הניצחונות המקסימאלי הוא שלושה. שאר הערכים בטבלה מציגים את **אחוז** ההצבעות שקיבל המודל לעומת המודלים האחרים. במידה והערך חיובי הדבר מצביע על כך כי היו יותר מסווגים אשר סיווגו את הרשתות האמפיריות למודל המופיע בשורה המתאימה מאשר למודל המופיע בעמודה המתאימה ולהיפך. הערכים הם בטווח -100%-100%. לכן, למשל, אם הערך המופיע הוא 60% המשמעות הינה ש-80% מהמסווגים סיווגו את הרשתות האמפיריות למודל המופיע בשורה ולא למודל המופיע בעמודה.

כפי שניתן לראות בטבלה, המודל "קישור מועדף" סטטי ממושקל ( WSPA ), שהינו הרחבה של המודל "קישור מועדף" סטטי אשר היה בעל ההתאמה הטובה ביותר במקרה הבינארי, הוא בעל ההתאמה הטובה ביותר גם במקרה זה. **אולם** בניגוד למקרה הבינארי בו ההתאמה הייתה חד משמעית בכל המסווגים, הפעם מדובר על התאמה פחות משמעותית במסווגים שבדקו את ההתאמה שלו לעומת המודלים WDPA ו-WIDPA.

) התאמה בכ-80% מהמקרים, במקום 100% במקרה של בדיקת הקישוריות בלבד).

בעקבות התוצאות הלא מספקות בוצע ניסיון למצוא מודל אחר שיתאים יותר במקרה הממושקל. הניסיון הראשון היה באמצעות שימוש במודל מס' 5WSDPA – , שתואר לעיל, אשר משלב שני מנגנונים ליצירת הקישורים ברשת. המניע לכך היה המחשבה, ששילוב המנגנונים של המודלים המובילים יביא ליצירת מודל בעל התאמה טובה יותר לנתוני הרשתות האמפיריות. אולם בדיקה חוזרת שנעשתה הראתה כי זה לא המקרה. למעשה כפי שניתן לראות בטבלה 8, המודל המשולב כשל באופן ברור והיה הגרוע ביותר מבין כל המודלים שנבחנו.

**טבלה 8 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים, כולל המודל המשולב.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **מודל/מודל** | **הצבעות** | **WDRG** | **WSDPA** | **WSPA** | **WDPA** | **WIDPA** |
| WDRG | 1/4 |  | 88.57% | -100.00% | -80.00% | -77.14% |
| WSDPA | 0/4 | -88.57% |  | -91.43% | -94.29% | -94.29% |
| **WSPA** | **4/4** | **100.00%** | **91.43%** |  | **62.86%** | **57.14%** |
| WDPA | 2/4 | 80.00% | 94.29% | -62.86% |  | -65.71% |
| WIDPA | 3/4 | 77.14% | 94.29% | -57.14% | 65.71% |  |

**טבלה 8:** טבלה זו מכילה בעמודה **הצבעות** את **מספר** המסווגים שסיווגו את הרשתות האמפיריות למודל המופיע באותה שורה. בסה"כ נבדקו חמישה מודלים.. כל מודל סווג באמצעות ארבעה מסווגים, ולכן מספר ההצבעות המקסימאלי הוא ארבע. משמעות הערכים בטבלה דומה למשמעותם בטבלה 7.

עקב הכישלון בשילוב המנגנונים נוסה כיוון חדש. כזכור מודל 2 ( WSPA ) הינו בעל פרמטרים ו-, שערכם המקורי היה 0.5. בניסיון זה בדקנו האם שינוי ערכם של פרמטרים אלו יגרום למודל WSPA להיות בעל התאמה טובה יותר לרשתות האמפיריות. בוצעו מספר ניסויים בהם שונה ערך הפרמטרים ו- בהדרגה מ-0.5 ל-0.6, 0.66 ו-0.7. תוצאות הניסויים מוצגות בטבלה 9.

**טבלה 9 - ריכוז תוצאות סיווג הרשתות האמפיריות למודלים, כולל שינויי ערכי הפרמטרים של מודל ה-WSPA.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **מודל/מודל** | **הצבעות** | **WDRG** | **WSPA** | **WSPA6** | **WSPA66** | **WSPA7** | **WDPA** | **WIDPA** |
| WDRG | 0/6 |  | -100.0% | -100.0% | -100.0% | -100.0% | -80.0% | -77.1% |
| WSPA | 3/6 | 100.0% |  | -100.0% | -100.0% | -100.0% | 62.9% | 57.1% |
| WSPA6 | 4/6 | 100.0% | 100.0% |  | -100.0% | -100.0% | 77.1% | 74.3% |
| WSPA66 | 5/6 | 100.0% | 100.0% | 100.0% |  | -100.0% | 82.9% | 82.9% |
| WSPA7 | 6/6 | 100.0% | 100.0% | 100.0% | 100.0% |  | 85.7% | 91.4% |
| WDPA | 1/6 | 80.0% | -62.9% | -77.1% | -82.9% | -85.7% |  | -65.7% |
| WIDPA | 2/6 | 77.1% | -57.1% | -74.3% | -82.9% | -91.4% | 65.7% |  |

**טבלה 9:** טבלה זו מכילה בעמודה **הצבעות** את **מספר** המסווגים שסיווגו את הרשת למודל המופיע באותה שורה. בסה"כ נבדקו שבעה מודלים. כל מודל סווג באמצעות שישה מסווגים, ולכן מספר ההצבעות המקסימאלי הוא שישה. משמעות הערכים בטבלה דומה למשמעותם בטבלה 7. המודלים WSPA6,WSPA66,WSPA7 הינם הווריאציות של המודל WSPA עם הפרמטרים השונים. (WSPA6 משמעותו שערך הפרמטרים שונה ל-0.6, וכן לגבי שאר הווריאציות ).

כפי שניתן לראות בטבלה, ניסויים אלו הראו כי שינוי ערך הפרמטרים אכן משפר את רמת ההתאמה וככל שערך הפרמטר עולה כך גדלה ההתאמה של המודל לרשתות האמפיריות. השוואת אחוז הניצחונות לתוצאות הבדיקה הראשונית מראה כי חל שיפור דרמטי של 11.4% ברמת ההתאמה לעומת מודל ה-WDPA ואף שיפור גבוה יותר של 17.15% ברמת ההתאמה לעומת מודל ה- WIDPA.

* בדיקת השיפור בתוצאות כתלות בפרמטרים בוצעה רק עד הערך 0.7. זאת הואיל ועבור ערך זה התקבלה תוצאה העומדת ביעד שהצבנו לעצמנו מראש שהינו התאמה של למעלה מ-90%. עם זאת, ייתכן כי כיוונון נוסף של הפרמטרים יאפשר להשיג התאמה טובה יותר.

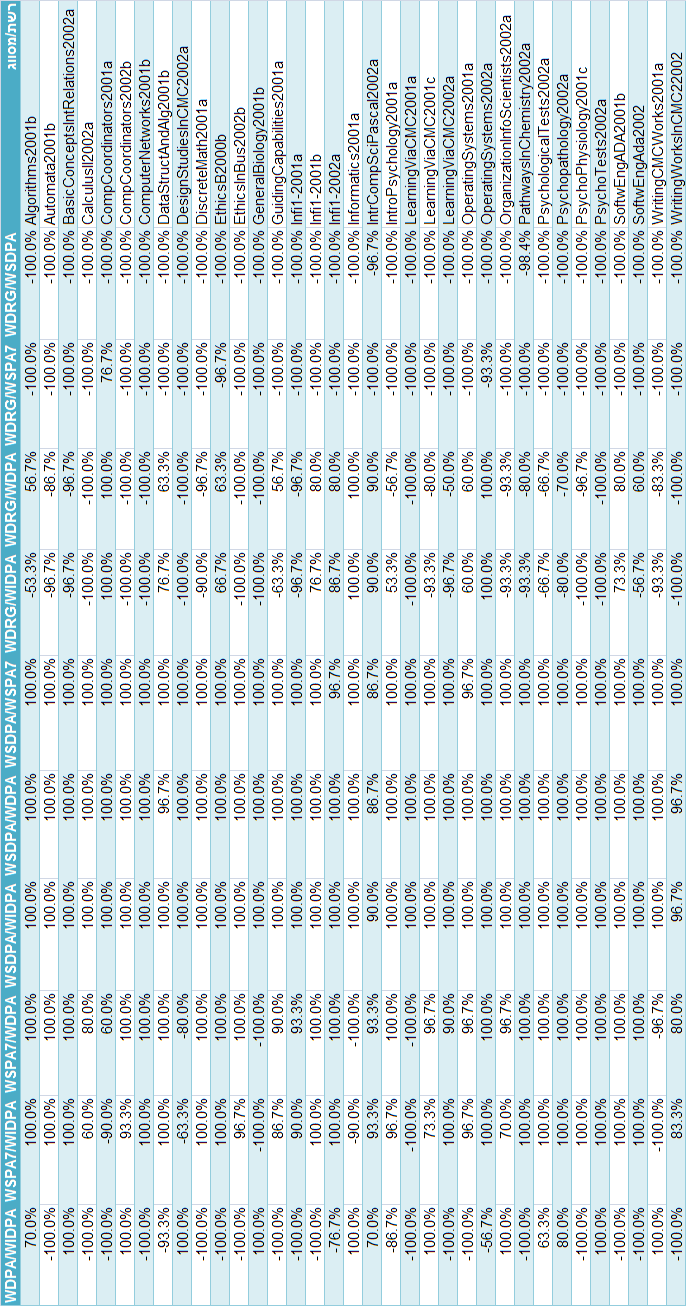
### בדיקת עמידות המסווגים

לאחר ההצלחה במציאת מודל בעל תוצאות יותר משמעותיות, בוצע שוב השלב של בדיקת עמידות המסווגים. בדומה למקרה הבינארי, עבור כל זוג מודלים נבחרו 30 תתי קבוצות בעלות 500 מאפיינים כל אחת ( כ-10% מכמות המאפיינים הכוללת ). עבור כל תת קבוצה נבנה מסווג מסוג SVM, ונבדקו תוצאות הסיווג שלו על הרשת האמפירית המתאימה.בתרשים 4 מוצגת היסטוגרמה המתארת את ערכי העמידות של ארבעה מן המסווגים. טבלה 10 מוצגים הערכים עליהם מתבססת ההיסטוגרמה ביתר פירוט.

( טבלה 10 מכילה את ערכי העמידות של כלל המסווגים, ולא רק של הארבעה המוצגים בהיסטוגרמה ).

**תרשים 4 - היסטוגרמת עמידות המסווגים עבור המקרה הממושקל**

**תרשים 4:** התרשים מכיל היסטוגרמה המציגה את הכמויות השונות של ערכי העמידות של ארבעה מן המסווגים שנבדקו. הציר האופקי הינו ערך העמידות והציר האנכי הינו כמות המסווגים מכל סוג ( מתוך 35 מסווגים ) שערך העמידות שלה היה בטווח. ***טבלה 10 – עמידות המסווגים עבור המקרה הממושקל***

****

**טבלה 10:** טבלה זו מכילה את פירוט תוצאות בדיקת העמידות של המסווגים השונים. הערך הרשום בתאים השונים הינו אחוז המסווגים אשר היו בעלי תוצאת סיווג זהה. ערך חיובי/שלילי בתא משמעותו כי הרשת סווגה אל המודל השמאלי/ימני בהתאמה כפי שמופיע בכותרת העמודה המתאימה.

כפי שניתן לראות, העמידות שגילו המסווגים הייתה גבוהה מאוד. עבור המסווגים שהוצגו בהיסטוגרמה למעלה מ-77% היו בעלי עמידות מושלמת, ועבור למעלה מ-90% הייתה עמידות של לפחות 90%. נתונים אלו דומים גם עבור כלל המסווגים: בכמעט 73% מהמקרים, המסווגים היו בעלי עמידות מושלמת ועבור למעלה מ-86% עמידות המסווג הייתה לכל הפחות 90%. עם זאת למרות העמידות הגבוהה של המסווגים, העמידות במקרה זה הייתה מעט נמוכה יותר מאשר במקרה הבינארי.

### בדיקת איכות תוצאות הסיווג

על מנת לוודא כי התוצאות שהתקבלו אכן בעלות מובהקות סטטיסטית, בוצעה בדיקה של מובהקות התוצאות בדומה לבדיקה שבוצעה בבדיקת המקרה הבינארי. במסגרת בדיקה זו, חושב עבור כל סוג מסווג, הקשר בין כמות הסיווגים השגויים של הרשתות השונות לבין ההסתברות של הסיווג. תוצאות הבדיקה עבור המסווגים בהם השתתף המודל המנצח ( WSPA7 ) מוצגות בתרשים מס' 5, ואילו בטבלה 11 מוצגים הנתונים עליהם מתבסס התרשים.

**תרשים 5- הקשר בין כמות טעויות הסיווג המירבית לבין הסתברות תוצאות הסיווג עבור המסווגים השונים במקרה הממושקל.**

**תרשים 5:** בתרשים זה ניתן לראות כיצד ככל שמגדילים את כמות טעויות סיווג האפשרית, כך ההסתברות של הסיווג הכולל הולכת וגדלה.

**טבלה 11: נתונים מפורטים של הסתברות תוצאות הסיווג בהתאם לכמות השגיאות המירבית.**

|  |  |  |  |
| --- | --- | --- | --- |
| **כמות שגיאות/מסווג** | **WDRG/WSPA7** | **WSPA7/WDPA** | **WSPA7/WIDPA** |
| 0 | 0.9993 | 0.1321 | 0.1440 |
| 1 | 1.0000 | 0.4625 | 0.5005 |
| 2 | 1.0000 | 0.7827 | 0.8223 |
| 3 | 1.0000 | 0.9422 | 0.9598 |
| 4 | 1.0000 | 0.9895 | 0.9939 |
| 5 | 1.0000 | 0.9987 | 0.9993 |
| 6 | 1.0000 | 0.9999 | 0.9999 |

**טבלה 11-**טבלה זו מציגה את הסתברות הסיווג המדויקת עבור כל מסווג בהתאם לכמות השגיאות המירבית. ניתן לראות כי ברוב המקרים הגענו להסתברות גבוהה מאוד של 0.999 תוך מספר שגיאות מצומצם ביותר.

חשוב להזכיר כי לצורך בדיקת המובהקות בוצע שימוש רק במסווגים שערך העמידות שלהם עבר את הסף של 90%. הגבלה זו גרמה לכך כי בחלק מהמקרים הבדיקה בוצעה על מספר נמוך יותר של תוצאות. טבלה מס' 12 מכילה פירוט של כמות המסווגים שהיו בשימוש בבדיקת המובהקות.

**טבלה מס' 12- כמות המסווגים שנבדקו במסגרת בדיקת המובהקות**

|  |  |
| --- | --- |
| **מסווג** | **כמות תוצאות** |
| WDRG/WSPA7 | 34 |
| WSPA7/WDPA | 27 |
| WSPA7/WIDPA | 25 |

**טבלה 12:** בטבלה ניתן לראות כי עמידות המסווגים הייתה טובה, ובמקרה הגרוע ביותר היה שימוש בלמעלה מ- 71% מהתוצאות ( 25 מתוך 35 ) למטרת בדיקת המובהקות. אמנם כפי שכבר צוין, בהשוואה למקרה הבינארי העמידות הייתה נמוכה יותר.

ניתן לראות כי מתקבלות תוצאות בעלות הסתברות גבוהה, עבור מספר שגיאות נמוך. למעשה, מספיק להניח כי למסווג היו לכל היותר 4 שגיאות בסיווג על מנת לקבל כי תוצאות הסיווג הינן בעלי הסתברות של למעלה מ-99.9%. כמות שגיאות זו שקולה לכך שבמקרה הגרוע ביותר, 84% מהרשתות סווגו נכונה ( במקרה של המסווג WSPA/WIDPA ) ועבור שאר המסווגים התוצאות אף טובות יותר.

לכן בדומה למקרה הבינארי, גם במקרה של בחינת הרשתות עם המשקלים אנו מגיעים למסקנה כי המודל הקרוב ביותר לרשתות האמפיריות הוא: "קישור מועדף" סטטי ממושקל ( WSPA ), אם כי ישנו שימוש בפרמטרים שונים במקצת.

# דיון ומסקנות

העובדה כי מודל הגרף האקראי (DRG) אינו מהווה מודל מתאים ליצירת רשתות לימוד מקוונות, לא מפתיעה. דבר זה דומה לתוצאות שהתקבלו עבור רשתות אמפיריות נוספות (Newman, 2003). כמו-כן עובדה זו הוכחה במחקרים קודמים על רשתות לימוד מקוונות אשר ניתחו מאפיינים ספציפיים של הרשתות (Aviv et al., 2007b). במחקר זה מחוזקת מסקנה זו באמצעות ניתוח ווקטור המאפיינים הרב מימדי של הרשתות.

גם העובדה שמודל ה"עולם הקטן" המכוון ( DSW ) אינו מהווה מודל מתאים ליצירת רשתות לימוד מקוונות, לא מפתיעה. אחד המאפיינים העיקריים של מודל זה הינו ערך גבוה של מקדם ה-Clustering. במקרה שלנו, משמעות הדבר היא שהמשתתפים ברשת נוטים להגיב בתוך קבוצות קטנות. דבר זה איננו מתרחש עבור רשתות הלימוד המקוונות. תוצאות אלו מתאימות למחקרים קודמים אשר בחנו את הנושא (Aviv et al., 2007a).

מודל ה"העתקה דינמית" ( DC ) מבוסס על לימוד של הקשרים ברשת. הלימוד הינו חלקי בלבד ומתבסס על גילוי השכנים של קודקודים אחרים. מודל זה מוצלח בתיאור רשתות ביולוגיות וטכניות ( רשתות ציטוטים ). אולם נראה כי מנגנון לימוד זה, אינו לוקח חלק בתהליך ההתפתחות של רשתות לימוד מקוונות. ניתן להסביר זאת בכך שברשתות הלימוד המקוונות המידע על כלל הרשת נגיש, בצורה יחסית נוחה, לכלל המשתתפים. לכן ההחלטות של המשתתפים האם ולמי להגיב מבוססות על **כלל** המידע על מבנה הרשת.

מודל ה"קישור מועדף" סטטי ( SPA ) מתאים כמעט באופן מושלם ליצירת הרשתות האמפיריות. מודל זה מניח כי ישנה, מראש, השמה לא אחידה של הסתברויות ליצירה וקבלת תגובות לכל אחד מהמשתתפים השייכים לקבוצת רשת הלימוד המקוונת. התוצאה היא התפלגות לא אחידה של ההסתברות ליצור או לקבל תגובות ברשת. לכן, התפקיד/העמדה הראשונית של המשתתפים ברשת הינו בעל חשיבות מרובה, כאשר משתתף מחליט האם ולמי להגיב. חשיבות זו אינה משתנה במהלך התפתחות הרשת אלא נשארת קבועה.

התנהגות זו הינה בניגוד למודל ה"קישור מועדף" דינמי (DPA). במודל זה המשתתפים מתעדכנים לגבי מצב הרשת באופן דינמי. בכל פעם שהם מחליטים להגיב, הם "מסתכלים" ובודקים את עוצמת המשיכה של המשתתפים השונים. ההחלטה על יצירת או קבלת התגובות מושפעת בהתאם, כאשר **ישנה עדיפות להגיב לבעלי "משיכה" גבוהה**. נראה כי תהליך למידה דינמי של מצב הרשת איננו מתרחש במקרה של רשתות לימוד מקוונות.

התוצאות עבור מודל ה"קישור מועדף" דינמי משופר (IDPA) טובות מעט יותר מאשר התוצאות עבור מודל ה"קישור מועדף" דינמי (DPA). כזכור, ההבדל בין שני המודלים הינו ביכולתו של המודל המשופר ליצור רשתות בעלות מספר רכיבי קשירות. השיפור בתוצאות בא לידי ביטוי בכך שיש צורך להניח קיום של יותר טעויות אפשריות בסיווג, על מנת שרמת המובהקות של ניצחון מודל הקישור מועדף סטטי (SPA) על מודל זה תהיה גבוהה כמו עבור שאר המקרים. דבר זה נובע, בעיקר מהעובדה שאיכות המסווגים עבור מודל זה הייתה נמוכה יותר מאשר המודלים של שאר המסווגים. מהמידע לגבי קישוריות הרשתות בלבד נראה כי התשובה לשאלה "האם למודל ישנה עדיפות לעומת המודל המקורי (DPA) בתיאור רשתות הלימוד המקוונות?" היא חיובית. אולם כאשר בוחנים את התוצאות של בדיקת עוצמת הקשרים ברשת ניתן לראות כי דבר זה אינו חד משמעי , כפי שיתואר בהמשך.

**כאשר עוברים לבחון את התוצאות של בחינת עוצמת הקשרים ברשת התמונה משתנה מעט.**

במקרה של מודל הגרף האקראי הממושקל (WDRG) עדיין ניתן לראות בבירור כי מודל זה אינו מהווה מועמד מתאים ליצירת רשתות הלימוד המקוונות בדומה לתוצאות בחינת הקישוריות.

אולם הפעם ישנה תחרות עזה יותר בין הגרסאות הממושקלות של המודלים לקישור מועדף דינמי (WDPA,WIDPA) לבין המודל לקישור מועדף סטטי (WSPA). אמנם עדיין המודל הסטטי מנצח, אך מדובר על ניצחון פחות משמעותי ( ניצחון בכ-80% מהמקרים ) שקיימת הסתברות גבוהה יותר שהוא עשוי להתרחש עקב טעויות בסיווג.

שיפור במצב מתקבל כתוצאה משינוי של הפרמטרים של המודל קישור מועדף סטטי. פרמטרים אלו שולטים על ההשמה של ההסתברויות ליצירה וקבלת תגובות לכל אחד מהמשתתפים השייכים לקבוצת רשת הלימוד המקוונת. הבדיקה שנערכה הראתה כי הגדלת ערכי הפרמטרים האלו מ-0.5 דרך 0.6 ו-0.66 ועד ל-0.7 מראה כי מתקבל שיפור ברמת ההתאמה של המודל. כאשר עבור השינוי ל-0.7 ( המודל WSPA7) ישנו שיפור משמעותי ברמת ההתאמה של המודל לרשתות האמפיריות. שינוי זה גורם לכך כי חלוקת ההסתברויות בין המשתתפים תהיה קיצונית יותר, ולמעשה מראה כי ההשפעה של התפקיד/עמדה של המשתתף ברשת הינה **משמעותית יותר** ממה שסברנו בתחילה בבחינת הקישוריות בלבד.

כמו-כן ניתן לראות כי ההבדלים שנצפו בבחינת הקישוריות בין המודל "קישור מועדף" דינמי ו"קישור מועדף" דינמי משופר מטשטשים וניתן אף לראות היפוך מגמה ככל שערכי הפרמטרים של מודל ה"קישור מועדף" סטטי הולכים וגדלים ( ראה טבלה 9 ). חיזוק לכך ניתן לראות גם באיכות הסיווג של המסווג בין שני מודלים אלו (WDPA/WIDPA ) אשר הינה כ-50% ( ראה טבלאות 6א' ו-6ב' ). כלומר, **אין כמעט הבדל** בין שני המודלים כאשר מתייחסים לעוצמת הקשרים ברשתות הלימוד המקוונות.

הקביעה מראש של הסתברויות התגובה של המשתתפים במודל המנצח – מודל ה"קישור מועדף" סטטי (SPA,WSPA7) - מושפעת, כפי הנראה, ממאפיינים חיצוניים של המשתתפים. למעשה במקרה מסוים (Aviv, Erlich, Ravid, & Geva, 2003) הראו כי בנייה מתוכננת של רשת לימוד מקוונת, הביאה לכך כי קבוצת סטודנטים הפכה לגורם מעורר ( גורם ליצירת תגובות ) ומגשר ברשת. כתוצאה מכך בניית הידע ברשת הגיעה לשלב מתקדם. ללא סטודנטים אלו, ייתכן כי מבנה הרשת היה מפוצל יותר והתהליך הלימודי היה פחות טוב, כפי שהיה ברשת לא מובנית שנבחנה. שינוי תפקידים זה גרם לכך כי למנחה של הקורס הייתה השפעה מועטה יותר על רשת הלימוד. השפעה זו התחלקה בין מהסטודנטים ובכך יצרה חלוקה לא אחידה של ההסתברות ליצור או לקבל תגובות ברשת.

# סיכום וכיוונים להמשך

המטרה של עבודת תזה זו הייתה למצוא את המנגנון השולט על יצירת רשתות לימוד מקוונות. בניגוד לעבודות קודמות, בעבודה זו זיהוי המודל לא התבסס על מספר מצומצם של מאפיינים, אלא נעשה שימוש בווקטור מאפיינים רב-מימדי. ווקטור זה הכיל מאפיינים של הרשתות שנוצרו באופן שיטתי ולא מוטה. מאפיינים אלו שימשו ליצירת מסווגים בעלי דיוק ועמידות גבוהים, בהם נעשה שימוש על מנת לסווג בין מספר מודלים ליצירת רשתות, אשר מיישמים מנגנונים שונים. הסיווג נעשה הן בהתבסס על הקישוריות ברשתות והן בהתבסס על עוצמת הקשרים. תוצאות הסיווג הראו כי ברמת מובהקות גבוהה המודל המתאים ביותר הינו מסוג "קישור מועדף" סטטי. מנגנון זה מקצה מראש הסתברויות ליצירה וקבלת תגובות למשתתפי הרשת באופן לא אחיד. **המשמעות המעשית של מנגנון זה הינה שהמעמד ההתחלתי של המשתתף ברשת משפיע באופן ניכר מאוד על כמות התגובות שמשתתף זה ייצור או יקבל**. מנחה של רשת לימוד מקוונת יכול להשתמש בתוצאה זו על מנת לתכנן את הרכב הרשת על מנת לקבל רמת תגובתיות רצויה.

עם זאת למחקר זה היו מספר מגבלות. המגבלה הברורה ביותר הינה בכמות המוגבלת של מודלים שנבחנו. אמנם המודלים שנבחנו ייצגו "משפחות" של מודלים המיישמים מנגנונים שונים, וההתאמה שקיבלנו כבר עכשיו הינה ברמה טובה מאוד, אך עדיין קיימת האפשרות כי קיים מודל אחר המתאר בצורה טובה יותר את הרשתות. עם זאת התוצאות שהתקבלו, נותנות בידינו את האפשרות לבחון בצורה יעילה יותר האם מודל חדש הוא בעל פוטנציאל התאמה גבוה יותר או לא, באמצעות התמקדות בהשוואה למודל הזוכה הנוכחי.

מגבלה נוספת במחקר טמונה בהנחה כי המודל המייצר את הרשתות מבוסס על מנגנון בודד. הנחה זו איננה בהכרח נכונה. אמנם הניסיון שעשינו לשלב מנגנונים, בבחינת עוצמת הקשרים ברשת ( מודל ה-WSPDA ), נכשל מבחינת התאמתו לנתוני הרשתות, אולם אין מן הנמנע שמודל אחר בעל שילוב מנגנונים שונה או בעל פרמטרים שונים יצליח יותר.

רובם המכריע של המודלים שבדקנו הראו רמה גבוהה מאוד של דיוק ועמידות. דבר זה מרמז על כך שההסתברויות המותנות של מגוון המאפיינים של המודלים השונים יהיו בעלי התנהגות דומה. כלומר, ערכי המאפיין ברשת האמפירית, יהיו בעלי ערך מקסימאלי עבור אותו מודל. דבר זה מעלה את האפשרות שאין צורך להשתמש בכל המאפיינים על מנת להבדיל בין המודלים, הואיל וכולם משפיעים על המסווג כך שיסווג את הרשת לאותו מודל. על מנת להעריך את רמת ההשפעה של כל מאפיין יש לחשב את ההסתברות המותנית של כל מאפיין לסיווג רשת למודל, בהינתן שלמאפיין זה יש ערך מסויים.

זיהוי המאפיינים הדומיננטיים יאפשר מספר דברים:

* בניית מסווגים בעלי איכות גבוהה המשתמשים במספר מאפיינים מצומצם יותר. ( עם זאת לא נראה שיש בכך תועלת מעשית ).
* בחינה בצורה מעמיקה יותר האם אכן יש אמת בהשערה שהועלתה לעיל כי שילוב של מספר מנגנונים הינו הגורם ליצירת הרשת. דבר זה ניתן לביצוע בדיקה האם המאפיינים הדומיננטיים, נוצרו ע"י מודל בודד או ע"י מספר מודלים. קיומם של מספר מודלים המייצרים את המאפיינים הדומיננטיים, יאפשר לנו לנסות וליצור מודל משולב אשר מייצר רשתות בעלות מאפיינים אלו.

כיוון מחקר נוסף הינו לבחון את התפתחות הרשת לאורך זמן. בבדיקה שנעשתה במחקר זה, בדקנו את הרשת הסופית, אולם ייתכן כי בדומה לקיומו של מעבר פאזה ברשתות אקראיות, קיים שלב דומה במהלך ההתפתחות של רשתות לימוד מקוונות. כלומר, ישנם מספר מנגנונים הנכנסים לפעולה בשלבי התפתחות שונים של הרשת. במידה ותופעה זו קיימת ייתכן כי ניתן לגלות אותה רק במידה והרשת קיימת לפרקי זמן ארוכים יותר, למשל אם מדובר על רשת של קורס שנתי ולא קורס סמסטריאלי.

# נספח א' - שיטת Middendorf לייצוג רשתות מורכבות במרחב תכונות רב מימדי

חלק ניכר מהעיסוק בתחום מחקר הרשתות הינו ניסיון לאפיין את מבנה הרשתות הנחקרות. איפיון הרשתות מבוצע באמצעות מאפיינים סטטיסטיים שונים אשר חלקם מתוארים בפרק 2. המאפיינים הנבדקים הינם מקומיים - מבוססים על קבוצת קודקודים, או גלובליים - מבוססים על כלל הרשת. בין המאפיינים המקומיים שנבדקו ניתן למנות את השוואת שכיחות של תתי גרפים ברשת הנתונה לעומת שכיחותם בגרפים בעלי התפלגות מסוימת. בין המאפיינים הגלובליים ניתן למנות את התפלגות דרגות הקודקודים בגרף, ועוד.

על מנת לבדוק מאפיינים מקומיים של רשתות פותח מושג ה-k-subgraph census. משמעותו בדיקת הכמויות של כל הווריאציות של גרפים בעלי k קודקודים בגרף, והשוואת ההתפלגות שלהן להתפלגות של תתי גרפים אלו בגרפים אקראיים. שיטה זו בעייתית משתי בחינות:

1. **חישובית:** ככל ש- גדל, כך גדלה מחלקת הגרפים השונים ( שאינם איזומורפיים). זוהי בעיה NP-Complete. כתוצאה מכך, השימוש בשיטה מצטמצם לערכי k קטנים (בד"כ ), אולם דבר זה גורם להטיה בתוצאות, הואיל ואין התחשבות במבנים גדולים יותר.
2. **חוסר יכולת הכללה:** שיטה זו מוגדרת אך ורק לגרפים "פשוטים" ( מכוונים ולא מכוונים ) ולא קיימת דרך ברורה להכללת השיטה עבור סוגים נוספים של גרפים, כמו גרפים ממושקלים.

על מנת להתגבר על בעיות אלו, נעשה לעיתים שימוש בהנחות בסיס. למשל במקום לבדוק את כל תתי הגרפים בגודל k, מניחים מראש שתתי גרפים מסוימים הוא החשובים ורק הם נבדקים. דוגמא לכך ניתן לראות במאמר המפורסם של Shen-Or et al (2002), שזיהה שלושה תתי גרפים בעלי משמעות ברשת הגנטית של E. Coli. בעיה נוספת המתגלה בבחינת מאמר זה היא בכך שלא הייתה שיטה אחידה לזיהוי תתי הגרפים הרלבנטיים באופן יעיל ולמעשה עבור כל תת גרף שזוהה "נתפר" אלגוריתם מיוחד למציאה שלו. Middendorf et al (2004) מציע דרך להתגבר על בעיות אלו באמצעות בנייה שיטתית של מאפיינים עבור הגרף.

## תיאור השיטה

השיטה מבוססת על כך שניתן לאפיין את הגרף באמצעות כמות המסלולים הפתוחים והסגורים הקיימים בו. על מנת לחשב את כמויות המסלולים האלו באופן יעיל, משתמשת השיטה באבחנה הבאה: תהי מטריצת השכנות של הגרף, כלומר אם קיימת קשת מקודקוד לקודקוד . הכפלת המטריצה בעצמה מייצרות מטריצה חדשה עבורה הינו מספר המסלולים באורך 2 מקודקוד לקודקוד ,. באופן כללי עבור הכפלה של המטריצה A בעצמה פעמים נקבל כי הערך במקום ה- שווה לכמות המסלולים באורך מקודקוד לקודקוד .

על מנת לחשב את כמות המסלולים הפתוחים והסגורים באורך כלשהו, אנו מגדירים שני היטלים של מטריצת הסמיכויות באופן הבא:

1. המטריצה מתקבלת באמצעות ההיטל האלכסוני של A, המוגדר באופן הבא:  
    ,
2. המטריצה שהינה המשלים של , כלומר .

בנוסף, על מנת להבדיל בין מסלולים בכיוונים שונים בגרף, משתמשים גם במטריצה המשוחלפת ( Transposed ) המסומנת ב- .

ארבעת מטריצות אלו מגדירות את האלפבית הבסיסי בן ארבעת האותיות . מהאותיות של אלפבית זה אנו יוצרים "מילים", שהינן רצף באורך כלשהו של אותיות האלפבית המייצגות רצף של הכפלות של המטריצות המתאימות. הואיל וחלק מהרצפים יוצרים את מטריצת ה-0, כגון: , שאינה תורמת לאפיון הגרף, לא מבוצע שימוש ישיר באלפבית הבסיסי אלא מורידים מספר רצפים מיותרים ובכך מקבלים "אלפבית מבצעי" ליצירת מילים המורכב מקבוצת הרצפים הבאה: .

תהי מילה המורכבת מרצף אותיות כלשהו של אותיות האלפבית המבצעי.על מנת שנוכל להשתמש במטריצה אותה מילה זו מייצגת, אנו מגדירים פונקציה , אשר מתרגמת את המטריצה המתקבלת *למספר ממשי.*

במאמר הוגדרו שתי פונקציות מתאימות שניתן להפעיל על המטריצה המתקבלת:

1. **Nnz**: מספר הזוגות הסדורים של קודקודים ברשת ביניהם יש מסלול.   
   פונקציה זו ניתנת לחישוב ע"י מניית האיברים במטריצה שערכם שונה מ-0.
2. **Sum**: כמות המסלולים השונים בין זוגות הקודקודים ברשת. יש לשים לב לכך כי בין כל זוג קודקודים יכול להיות יותר ממסלול אחד.   
   פונקציה זו ניתנת לחישוב ע"י סכימת איברי המטריצה.

כמו-כן במאמר בוצעה אבחנה בין מסלולים סגורים ומסלולים פתוחים, ולכן באופן מעשי בוצע שם שימוש בקבוצת הפונקציות הבאות: . מייצגות את הפעלת הפונקציות Nnz ו–Sum על המסלולים הפתוחים, כלומר על איברי המטריצה שאינם שייכים לאלכסון. מייצגות את הפעלת הפונקציות Nnz ו–Sum על המסלולים הסגורים, כלומר כלומר על איברי האלכסון של המטריצה. אמנם הערכים השונים אינם בלתי תלויים, למשל אולם במאמר נימקו את השימוש בהם בכך שייתכן שפרמטר מורכב יוכל לתת אבחנה טובה יותר מהפרמטרים הבסיסיים.

לאחר הגדרת האלפבית והפונקציות, מייצרים באופן שיטתי את כל הרצפים עד לאורך מסוים ( תיאורטית אין הגבלה על האורך ), ומחשבים את ערך הפונקציות עבור המילים שהתקבלו. כל אחד מערכים אלו מהווה מימד בווקטור המאפיינים הרב-מימדי המייצג את הגרף.

## יתרונות השיטה

לשיטה זו ישנן שלושה יתרונות:

1. השיטה יעילה יותר מבחינה חישובית: שיטת ה-k-census graph הינה בעלת סיבוכיות של NP-Complete, אשר תלויה במאפייני הגרף כגון הצפיפות והתפלגות דרגות הקודקודים. בשיטה זו הסיבוכיות חסומה ע"י הביטוי הבא: , כאשר N הינו מספר הקודקודים בגרף, הינו מספר האותיות (מהאלפבית הבסיסי) של המילה ה-, והסכימה מבוצעת על כל המילים שיוצרו. אמנם הסיבוכיות גדלה אקספוננציאלית כתלות במספר האותיות, אולם הביטוי האקספוננציאלי אינו תלוי במאפיינים של הגרף. לכן שיטה זו עדיפה כאשר מדובר בגרפים צפופים או "חסרי מידה" ( Scale Free ). כתוצאה מהיעילות החישובית ( היחסית ) של השיטה ניתן למצוא מאפיינים בולטים ברשת לעומת אוסף אקראי של רשתות , מבלי להניח מראש התפלגות כלשהי של ערכי המאפיין בעולם, כפי שהיה מקובל.
2. השיטה ניתנת להכללה עבור גרפים מסוגים נוספים, כגון: גרפים ממושקלים.
3. שיטה זו מאפשרת לייצר באופן שיטתי ולא מוטה מאפיינים של הרשת אשר עשויים לאפשר מציאת תכונות חדשות המייחדות את הרשת.

## חסרונות השיטה

1. שיטה זו אינה מאפשרת לבדוק את כל התכונות הסטטיסטיות של הרשתות באופן ישיר. למשל חישוב מקדם האשכול של הגרף לא ניתן לביצוע באופן פשוט באמצעות השיטה הואיל והוא מחייב חישוב עבור כל קודקוד בנפרד.

# נספח ב' - תיאור חבילת התוכנה ששימשה לביצוע הבדיקות

במסגרת עבודת תזה זו בוצע שימוש בחבילת תוכנה אשר שימשה לביצוע הבדיקות.

חבילת התוכנה פותחה באמצעות Visual Studio 2008 ( Express Edition ).

חבילת התוכנה מורכבת משלושה יישומים:

1. **מחולל המאפיינים - AttributesCreator**

יישום זה אחראי על יצירת ווקטורי המאפיינים של רשתות על פי שיטת Middendorf המתוארת בנספח א'.

1. **מסווג הרשתות**

יישום זה אחראי על סיווג הרשתות האמפיריות החל משלב בניית המסווגים בין המודלים השונים, דרך בדיקת הדיוק והעמידות של המסווגים וכלה בסיווג הרשתות האמפיריות. דבר זה מתבצע באמצעות מימוש חופשי (freeware) בשפת C של מסווג מסוג SVM בשם SVM-Light (Joachims, 1999)

1. **בודק המובהקות**

יישום זה אחראי לבדוק מהי רמת המובהקות של תוצאות הסיווג שהתקבלו ע"י מסווג הרשתות עבור רמות שונות של שגיאות בסיווג.

## ב.1. מחולל המאפיינים ( Attribute Creator )

### מבנה

המבנה של מחולל המאפיינים מורכב מן המחלקות הבאות:

* **המחלקה Model**

מחלקה זו הינה מחלקת האב ממנה יורשים כל המחלקות הממשות מודלים ליצירת רשתות.

**המחלקה מכילה את האיברים הבאים:**

1. **mModelName**: מחרוזת המתארת את שם המודל.
2. **mRand**: אובייקט מסוג Random המשמש ליצירת ערכים אקראיים שונים. אובייקט זה משמש למימוש הבחירות האקראיות המבוצעות במודלים השונים, בהתאם להגדרת המודל.
3. **mMatrices**: רשימה המכילה את הרשתות שנוצרו ע"י המודלים השונים. הרשימה מכילה את הרשתות כשהם מיוצגות ע"י מטריצת סמיכויות דלילה. ( SparseMatrix ).

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: משמש לאתחול המשתנים השונים של המחלקה.
2. **Create**: זוהי שיטה וירטואלית. שיטה זו ממומשת ע"י המחלקות היורשות מהמחלקה Model. השיטה אחראית על היצירה בפועל של הרשתות. השיטה מקבלת כפרמטר את כמות הרשתות שיש לייצר.
3. **CreateVectors**: שיטה זו אחראית על הפיכת מטריצות הסמיכויות של הרשתות שנוצרו ע"י המודלים השונים לווקטורי מאפיינים בשיטת Middendorf. לשיטה זו מועברים מספר פרמטרים:
   1. **maxLength**: אורך המילים המרבי שיש לייצר.
   2. **path**: מיקום שמירת התוצאות.
   3. **normalize**: האם לנרמל את התוצאות.
   4. **useWeights**: האם להתייחס למשקלים של הקשרים ברשת.
   5. **type**: באיזו שיטה לייצר את המאפיינים. פרמטר זה לא בשימוש והוא מיועד להרחבות עתידיות של היישום.
4. **WriteMatrices**: שיטה זו משמשת לכתיבת מטריצות הסמיכויות של הרשתות שנוצרו ע"י המודל לתיקיה המועברת כפרמטר לשיטה.
5. **GetMatrices**: שיטה זו משמשת להחזרת רשימת מטריצות הסמיכויות של הרשתות שנוצרו ע"י המודל.
6. **Normalize**: זוהי שיטה פנימית (private). שיטה זו מנרמלת ווקטור המייצג את המאפיינים של רשת, השיטה מופעלת ע"י השיטה CreateVectors במידה ויש צורך בכך. בפועל לאחר ביצוע ניסויים עם ובלי שימוש בשיטה הוחלט לוותר על השימוש בה.

* **המחלקה Matrix**

מחלקה זו משמשת לייצוג מטריצת הסמיכויות של רשת. המחלקה מייצגת את הרשת באמצעות מטריצה ריבועית צפופה (dense ). המחלקה מכילה שיטות שונות לחישוב תכונות שונות של הרשת וכן שיטות לביצוע פעולות על המטריצה, הנדרשות לחישוב מאפייני הווקטור .

**המחלקה מכילה את האיברים הבאים:**

1. **mRows**: כמות השורות במטריצה.
2. **mColumns**: כמות העמודות במטריצה.
3. **mMatrix**: מערך דו-מימדי המכיל את ערכי המטריצה.

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: השיטה מקבלת את כמות השורות והעמודות ומייצרת מטריצה בגודל המתאים אשר ערכיה מאותחלים ל-0.
2. **בנאי**: השיטה מקבלת את מיקומו של קובץ המכיל מטריצה, ואת התו המפריד בין הערכים ומייצרת מטריצה על סמך נתוני הקובץ.
3. **Read**: השיטה מקבלת את מיקומו של קובץ המכיל מטריצה, ואת התו המפריד בין הערכים ומייצרת מטריצה על סמך נתוני הקובץ.
4. **Write**: השיטה מקבלת את מיקומו של קובץ יעד, ואת התו המפריד בין הערכים וכותבת את המטריצה הקיימת לתוך הקובץ.
5. **Report**: השיטה כותבת נתונים על המטריצה. הנתונים נכתבים לתוך אובייקט מסוג TextWriter המועבר כפרמטר. ניתן לשלוט על רמת פירוט הנתונים באמצעות שני פרמטרים נוספים הקובעים האם לכתוב רק נתונים סיכומיים, והאם לכתוב את היסטוגרמת התפלגות דרגות הקודקודים.
6. **Clustering**: השיטה מחזירה את ערך מקדם ה-Clustering של המטריצה.
7. **SmallWorld**: השיטה מחזירה את ערך ה-SmallWorld של המטריצה.
8. **NodesInOutDegrees**: השיטה כותבת לתוך אובייקט מסוג TextWriter המועבר כפרמטר את היסטוגרמת התפלגות דרגות הכניסה והיציאה של הקודקודים.
9. **RunBFS**: זוהי שיטה פרטית הנמצאת בשימוש ע"י השיטה SmallWorld. השיטה משתמשת באלגוריתם BFS למטרת חישוב המרחקים מקודקוד מסויים לכל שאר הקודקודים ברשת, כשלב בחישוב ערך ה-SmallWorld.
10. **Get**: שיטה זו מחזירה את הערך של איבר מסויים במטריצה.
11. **Set**: שיטה זו קובעת את הערך של איבר מסויים במטריצה.
12. **Print**: שיטה זו כותבת את המטריצה לתוך TextWriter המועבר כפרמטר.
13. **Convert2Sparse**: שיטה זו ממירה את הייצוג של המטריצה לייצוג של מטריצה דלילה (SparseMatrix).
14. **Copy**: שיטה זו מחזירה העתק של המטריצה.
15. **Transpose**: שיטה זו מחזירה את המטריצה המשוחלפת (Transpose).
16. **Diagonal**: שיטה זו מחזירה את האלכסון של המטריצה.
17. **UnDiagnoal**: שיטה זו מחזירה את המשלים לאלכסון של המטריצה.
18. **Mult**: שיטה זו מכפילה את המטריצה במטריצה אחרת המתקבלת כפרמטר ומחזירה את תוצאת המכפלה.
19. **Edges**: שיטה זו מחזירה את כמות הקשתות ברשת.
20. **EdgesWeighted**: שיטה זו מחזירה את סכום משקלי הקשתות ברשת.
21. **AverageDegree**: שיטה זו מחזירה את הדרגה הממוצעת ברשת.
22. **AverageWeightedDegree**: שיטה זו מחזירה את המשקל הממוצע ברשת ממושקלת.
23. **InDegreeHist**: שיטה זו מחזירה את היסטוגרמת דרגות הכניסה של הקודקודים ברשת.
24. **OutDegreeHist**: שיטה זו מחזירה את היסטוגרמת דרגות היציאה של הקודקודים ברשת.
25. **Sum**: שיטה זו מחזירה את סכום איברי המטריצה.
26. **SumD**: שיטה זו מחזירה את סכום איברי האלכסון במטריצה.
27. **SumU**: שיטה זו מחזירה את סכום איברי המשלים לאלכסון במטריצה.
28. **Nnz**: שיטה זו מחזירה את כמות האיברים השונים מאפס במטריצה.
29. **NnzD**: שיטה זו מחזירה את כמות האיברים השונים מאפס באלכסון המטריצה.
30. **NnzU**: שיטה זו מחזירה את כמות האיברים השונים מאפס במשלים לאלכסון במטריצה.
31. **Rows**: שיטה זו מחזירה את כמות השורות במטריצה.
32. **Columns**: שיטה זו מחזירה את כמות העמודות במטריצה.
33. **Convert2Binary**: שיטה זו ממירה את המטריצה הניתנת כפרמטר למטריצה בינארית (מטריצה שערכי איבריה 0 או 1). השיטה נועדה להמרת מטריצות ממושקלות טרם חישוב ווקטור המאפיינים, במקרה שיש צורך לחשב את המאפיינים של הקשרים ברשת בלבד, ולא את המאפיינים של עוצמת הקשרים.

* **המחלקה SparseMatrix**

מחלקה זו משמשת לייצוג רשת באמצעות מטריצת סמיכויות דלילה, כלומר מטריצה בה נשמרים רק הערכים השונים מאפס. מחלקה זו נועדה לייעל את ייצור ווקטורי המאפיינים. הואיל וייצור ווקטורי המאפיינים מחייב פעולות רבות מאוד של הכפלת מטריצות, כאשר ישנו שימוש במטריצות דלילות, פעולות אלו מבוצעות באופן יעיל יותר מאשר ע"י שימוש בייצוג רגיל של המטריצה. עם זאת עדיין ישנו שימוש במחלקה Matrix, לצורך חישוב ערכים שונים אשר לא מבוצעים בצורה אינטנסיבית מחד, ומאידך נוחים יותר לביצוע בצורת הייצוג הרגילה של המטריצה.

**המחלקה מכילה את האיברים הבאים:**

1. **mRowsAmount**: כמות השורות במטריצה.
2. **mColumnsAmount**: כמות העמודות במטריצה.
3. **mRows**: רשימה המכילה עבור כל שורה מיפוי בין העמודה לבין הערך.
4. **mColumns**: רשימה המכילה עבור כל עמודה מיפוי בין המיקום בשורה לבין הערך.

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: השיטה מקבלת את כמות השורות והעמודות ומייצרת מטריצה בגודל המתאים ללא ערכים.
2. **בנאי**: השיטה מקבלת את מיקומו של קובץ המכיל מטריצה, ואת התו המפריד בין הערכים ומייצרת מטריצה על סמך נתוני הקובץ.
3. **InitRowsAndColumns**: השיטה מאתחלת את מבני הנתונים של המחלקה.
4. **Read**: השיטה מקבלת את מיקומו של קובץ המכיל מטריצה, ואת התו המפריד בין הערכים ומייצרת מטריצה על סמך נתוני הקובץ.
5. **Write**: השיטה מקבלת את מיקומו של קובץ יעד, ואת התו המפריד בין הערכים וכותבת את המטריצה הקיימת לתוך הקובץ. ( המטריצה נכתבת כמטריצה רגילה, ולא כמטריצה דלילה ).
6. **Report**: השיטה כותבת נתונים על המטריצה. הנתונים נכתבים לתוך TextWriter המועבר כפרמטר. ניתן לשלוט על פירוט הנתונים באמצעות שני פרמטרים נוספים הקובעים האם לכתוב רק נתונים סיכומיים, והאם לכתוב את היסטוגרמת התפלגות דרגות הקודקודים.
7. **Get**: שיטה זו מחזירה את הערך של איבר מסוים במטריצה.
8. **Set**: שיטה זו קובעת את הערך של איבר מסוים במטריצה.
9. **Convert2Matrix**: שיטה זו ממירה את הייצוג של המטריצה לייצוג של מטריצה רגילה (Matrix).
10. **Transpose**: שיטה זו מחזירה את המטריצה המשוחלפת (Transpose).
11. **Diagonal**: שיטה זו מחזירה את האלכסון של המטריצה.
12. **UnDiagnoal**: שיטה זו מחזירה את המשלים לאלכסון של המטריצה.
13. **Mult**: שיטה זו מכפילה את המטריצה במטריצה אחרת המתקבלת כפרמטר ומחזירה את תוצאת המכפלה.
14. **Sum**: שיטה זו מחזירה את סכום איברי המטריצה.
15. **SumD**: שיטה זו מחזירה את סכום איברי האלכסון במטריצה.
16. **SumU**: שיטה זו מחזירה את סכום איברי המשלים לאלכסון במטריצה.
17. **Nnz**: שיטה זו מחזירה את כמות האיברים השונים מאפס במטריצה.
18. **NnzD**: שיטה זו מחזירה את כמות האיברים השונים מאפס באלכסון המטריצה.
19. **NnzU**: שיטה זו מחזירה את כמות האיברים השונים מאפס במשלים לאלכסון במטריצה.
20. **Rows**: שיטה זו מחזירה את כמות השורות במטריצה.
21. **Columns**: שיטה זו מחזירה את כמות העמודות במטריצה.
22. **Convert2Binary**: שיטה זו ממירה את המטריצה הניתנת כפרמטר למטריצה בינארית (מטריצה שערכי איבריה 0 או 1). השיטה נועדה להמרת מטריצות ממושקלות טרם חישוב ווקטור המאפיינים, במקרה שרוצים לחשב את המאפיינים של הקשרים ברשת בלבד, ולא את המאפיינים של עוצמת הקשרים.

* **המחלקה NodeInfo**

מחלקה זו הינה למעשה מבנה נתונים המשמש להחזקת מידע על קודקוד ברשת באופן מרוכז. במידע זה מבוצע שימוש בחלק מן המודלים לצורך ההחלטה אלו קודקודים יקושרו זה לזה.

**המחלקה מכילה את האיברים הבאים:**

1. **mNodeIndex**: המספר הסידורי של הקודקוד ברשת.
2. **mNodeInDegree**: דרגת הכניסה של הקודקוד.
3. **mNodeInWeight**: המשקל היחסי של דרגת הכניסה של הקודקוד מסכום דרגות הכניסה ברשת.
4. **mNodeOutDegree**: דרגת היציאה של הקודקוד.
5. **mNodeOutWeight**: המשקל היחסי של דרגת היציאה של הקודקוד מסכום דרגות היציאה ברשת.

* עבור כל משתנה, קיימת במחלקה שיטה לקביעת/קבלת ערכו.
* **המחלקה ModelVectors**

תפקידה של מחלקה זו הינו להחזיק את רשימת המאפיינים של הרשתות אשר חוללו במהלך הרצת היישום. מחלקה זו ממומשת ע"י התבנית של Singleton. ( קיים רק מופע יחיד של המחלקה, הנגיש לכלל חלקי היישום ).

**המחלקה מכילה את האיברים הבאים:**

1. **mAttributes**: רשימה המכילה את המאפיינים של הווקטורים הנוצרים.

**המחלקה מכילה את השיטות הבאות:**

1. **Attributes**: שיטה זו מאפשרת גישה לרשימת המאפיינים. השימוש בה הינו בהוספת מאפיינים לרשימה תוך כדי החישוב שלהם.
2. **writeAttributes**: שיטה זו כותבת את רשימת המאפיינים לקובץ. מיקום הקובץ ניתן כפרמטר של השיטה.

* **המחלקה TrueNetModelWrapper**

מחלקה זו יורשת מהמחלקה Model. המחלקה משמשת לעטיפה (Wrapping) של רשת קיימת, כך שיהיה ניתן להפעיל עליה את השיטות של המחלקה Model. זאת על מנת לאפשר את ייצור ווקטור המאפיינים עבור הרשת. במחלקה זו נעשה שימוש על מנת לייצר את ווקטור המאפיינים של הרשתות האמפיריות.

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: הבנאי מקבל רשת קיימת ( בצורת SparseMatrix ) ומוסיף אותה לרשימת הרשתות, המוגדרת במחלקה Model, ממנה יורשת המחלקה.

* **המחלקה VectorGenerator**

מחלקה זו אחראית על יצירת ווקטור התכונות של הרשתות.

**המחלקה מכילה את האיברים הבאים:**

1. **mOrigMatrix**: משתנה זה מחזיק את המטריצה המקורית עליה יש לבצע את כל הפעולות.
2. **mWordMatrices**: משתנה זה הינו מסוג מיפוי בין "מילה" לבין המטריצה שהיא מייצגת. משתנה זה מיועד לייעל את חישוב המילים ע"י שמירת תוצאות ביניים שחושבו בשלבים מוקדמים יותר.

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: מאתחל את המשתנים של המחלקה.
2. **Create**: שיטה זו מקבלת את אורך המילים המרבי ואת סוג המאפיינים שיש ליצור, ומייצרת את ווקטור המאפיינים של הרשת. נכון לעכשיו השיטה תומכת רק ביצירת המאפיינים לפי שיטת Middendorf.
3. **EvaluateMiddendorf**: שיטה זו מקבלת רשימת "מילים" ( רצף של "אותיות", כאשר כל אות מייצגת פעולה על המטריצה של הרשת ), ומפעילה אותה על המטריצה של הרשת המקורית, על מנת לקבל את מטריצת התוצאה. לאחר קבלת מטריצת התוצאה, מופעלות עליה פונקציות שונות על מנת לייצר את הערכים של המאפיינים השונים. ה"אותיות" והפונקציות הן ע"פ השיטה של Middendorf et al ( 2004 ( המתוארת בנספח א'.
4. **GenerateMiddendorf**: שיטה זו מקבלת את האורך המרבי של ה"מילים" שיש לייצר ומייצרת אותם. ייצור ה"מילים" מבוצע באופן רקורסיבי.
5. **ApplyOperation**: שיטה זו מקבלת מטריצה ואת הפעולה שיש לבצע עליה, ומחזירה את מטריצת התוצאה.
6. **GetMatrixByOperation**: שיטה זו ממירה את המטריצה המקורית בהתאם לפעולה שיש לבצע בהתאם לאלפבית המבצעי. כלומר, היא ממירה את המטריצה המקורית לאחת המטריצות הבאות:
   1. המטריצה המשוחלפת – T.
   2. המטריצה האלכסונית – D.
   3. המטריצה המשלימה למטריצה האלכסונית – U.
   4. המטריצה המשולבת –AUT, כאשר A היא המטריצה המקורית. ( בקוד מסומנת מטריצה זו ב-C, על מנת שיהיה ניתן לייצג את כל אחת מהפעולות באמצעות תו בודד ).

* **מחלקות ליצירת רשתות על פי מודלים**

מלבד מחלקות אלו היישום מכיל עוד מחלקות רבות היורשות מהמחלקה Model, אשר כל אחת מהן מסוגלת לייצר רשתות על פי מודל ליצירת רשתות מסוים. מחלקות אלו כוללות שיטה סטטית בשם Run אשר אחראית לייצר את ווקטורי המאפיינים של המודל.

**לשיטה זו ישנם מספר פרמטרים:**

1. nExamplesAmount: כמות הרשתות שיש לייצר באמצעות המודל.
2. nMaxWordLength: האורך המרבי של מילים שיש ליצור על מנת לייצר את ווקטור המאפיינים.
3. ResultDir: התיקיה אליה יישמרו התוצאות.
4. bNormalize: האם לנרמל את ווקטורי המאפיינים.
5. bWriteMatrices: האם לשמור את הרשתות שנוצרו.
6. NodesAmount: כמות הקודקודים ברשת אשר המודל מנסה לחקות.
7. EdgesAmount: כמות הקשתות ברשת אשר המודל מנסה לחקות.
8. פרמטרים נוספים הנדרשים עבור המודלים השונים.

**שיטה זו פועלת באופן הבא:**

1. מתבצעת בדיקה האם קובץ היעד המכיל את ווקטורי המאפיינים של המודל קיים. במידה והקובץ כבר קיים, השיטה מסיימת את פעולתה מבלי לעשות דבר.
2. מיוצר מופע של המודל הרלבנטי, עם הפרמטרים המתאימים.
3. מתבצעת קריאה לשיטה Create של מופע המודל.
4. במידה ובקריאה לשיטה Run, הועבר פרמטר המבקש לשמור את הרשתות שנוצרו, אזי הן נשמרות לספרייה ResultDir.
5. מתבצעת קריאה לשיטה CreateVectors של מופע המודל ,אשר יוצרת את ווקטורי המאפיינים ושומרת אותו לתיקיה ResultDir.

* **המחלקה Program**

המחלקה Program אחראית על הפעולה שייבצע היישום.

**המחלקה מכילה את השיטות הבאות:**

1. **Main**: קריאת הפרמטרים איתם הורץ היישום והחלטה על ביצוע הפעולה המתאימה.
2. **CreateTrueNetVector**: שיטה זו מייצרת את ווקטורי המאפיינים של רשת אמפירית על סמך קובץ המכיל את מטריצת השכנות של הרשת. ווקטור זה נדרש על מנת לסווג את הרשת האמפירית למודלים באמצעות המסווגים שנבנו.
3. **CreateModelsVectors**: שיטה זו מייצרת את ווקטורי המאפיינים באמצעות המודלים ליצירת רשתות מכוונות.
4. **CreateModelsWeightedVectors**: שיטה זו מייצרת את ווקטורי המאפיינים באמצעות המודלים ליצירת רשתות מכוונות ממושקלות.

### הרצת מחולל המאפיינים

הרצת היישום ניתנת לביצוע באופן הבא:

**AttributesCreator**  -a <TR,TW,D,W> -s <networks dir> -d <destination dir>

-l <max word length> -n <0,1> -e <vectors to create> -w <0,1>

**יש לשים לב לנקודות הבאות:**

* ההרצה משתמשת בזוגות של פרמטרים: הראשון מצביע על סוג הפרמטר והשני על הערך שלו.
* אין חשיבות לסדר הזוגות של הפרמטרים.
* יש להוסיף רווח בין סוג הפרמטר לבין הערך שלו.
* אין חובה לרשום את כל הפרמטרים. במידה ופרמטר לא נרשם יבוצע שימוש בערך ברירת המחדל שלו. (את ערכי ברירת המחדל ניתן לראות בקוד המחלקה Program).

**משמעות הפרמטרים היא:**

1. **-a**: סוג הפעולה שיש לבצע, כאשר האפשרויות הן:
   1. **TR**: ייצור ווקטורי המאפיינים של הרשתות האמפיריות, תוך התייחסות אליהם כאל רשתות לא ממושקלות.
   2. **TW**: ייצור ווקטורי המאפיינים של הרשתות האמפיריות, תוך התייחסות אליהם כאל רשתות ממושקלות.
   3. **D**: ייצור ווקטורי המאפיינים של המודלים המכוונים הלא ממושקלים.
   4. **W**: ייצור ווקטורי המאפיינים של המודלים הממושקלים (המכוונים).
2. **-s**: מיקום קבצי הרשתות האמפיריות.
3. **-d**: תיקית היעד לכתיבת התוצאות.
4. **-l**: אורך ה"מילים" המרבי על פיהן ייווצרו המאפיינים.
5. **-n**: האם לנרמל את ווקטורי המאפיינים (0- לא לנרמל, 1- לנרמל).
6. **-e**: כמות הווקטורים שיש לייצר עבור כל מודל (במידה ובוחרים את האפשרויות T או D עבור הפרמטר –a).
7. **-w**: האם לשמור את הרשתות שנוצרו (0-לא לשמור, 1- לשמור).

**דוגמא:**

הפקודה הבאה תייצר 500 ווקטורי מאפיינים לא מנורמלים המבוססים על המילים עד אורך 5 עבור כל מודל ממושקל. הווקטורים ייווצרו על סמך קבצי הרשתות הנמצאות בספריה C:\Networks. התוצאות יישמרו ( יחד עם הרשתות עצמן ) בספריה C:\Results.

**AttributeCreator** –a W –s c:\Networks –d c:\Results –l 5 –n 0 –e 500 –w 1

### תיאור קבצי הרשתות

תיקיית הרשתות מכילה קבצים המכילים את מטריצת השכנות של הרשתות אותן ברצוננו לבדוק.

כל קובץ של רשת הינו קובץ טקסטואלי בעל המבנה הבא:

* 1. הקובץ מכיל מספר שורות, כאשר כל שורה מייצגת את השכנים של קודקוד מסויים ברשת.
  2. כל שורה מכילה מספר עמודות. מספר העמודות חייב להיות זהה למספר השורות. בקובץ.
  3. בכל עמודה מופיע מספר חיובי שלם. מספר זה מייצג את עוצמת הקשר בין הקודקוד המיוצג ע"י השורה לקודקוד המיוצג ע"י העמודה.
  4. בין העמודות מפריד תו רווח בודד.
  5. סדר הקודקודים בעמודות וסדר הקודקודים בשורות חייב להיות זהה.

**דוגמא:**

השורות הבאות מייצגות מטריצת שכנות של רשת מכוונת המכילה קליקה בגודל 4 ( רשת בעלת ארבעה קודקודים שכל אחד מהם מקושר לכל האחרים )

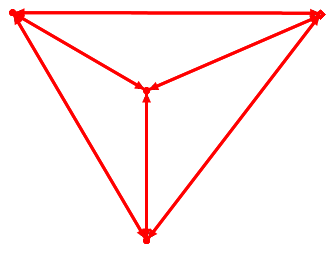
0 1 1 1

1 0 1 1

1 1 0 1

1 1 1 0

תיאור גרפי של הרשת



### תיאור התוצאות

* במקרה של יצירת ווקטורי המאפיינים של הרשתות האמפיריות, נוצרים בתוך תיקיית התוצאות קבצים נפרדים עבור כל רשת אמפירית. כל קובץ שכזה מכיל את ווקטור המאפיינים של הרשת הרלבנטית.
* במקרה של יצירת ווקטורי המאפיינים של המודלים, עבור כל רשת אמפירית נוצרת תחת תיקיית התוצאות, תיקיה ששמה כשם הרשת האמפירית. לתיקיה זו נכתבים הקבצים המכילים את ווקטורי המאפיינים של כל מודל, כאשר שם הקובץ מכיל את שם המודל הרלבנטי. כמו-כן נוצרת תחת תיקיה זו תיקיה המכילה את הרשתות שנוצרו ( במידה ובחרנו לשמור אותן ). למשל במידה והשתמשנו באפשרות זו על מנת ליצור ווקטור מאפיינים עבור רשת בשם N על סמך המודל שנקרא DRG, אזי תחת תיקיית התוצאות תיווצר תיקיה בשם N ובתוכה יהיה קובץ בשם DRG.vectors.
* קובץ ווקטור המאפיינים, הינו קובץ טקסטואלי בעל המבנה הבא:

1. בכל שורה יש ווקטור נפרד.
2. הווקטור מיוצג בצורה דלילה, כלומר מופיעים רק האיברים של הווקטור השונים מאפס והאינדקס שלהם.
3. האיברים של הווקטור מסודרים לפי האינדקס שלהם בסדר עולה.

לדוגמא: 1:216 2:11 3:205 4:237 5:21 6:216 7:104

## ב.2. מסווג הרשתות ( Networks Classifier )

### מבנה

המבנה של מסווג הרשתות מורכב מן המחלקות הבאות:

* **המחלקה Handler**

מחלקה זו היא מחלקת אב, ממנה יורשות מספר מחלקות. המחלקה מכילה איברים ושיטות הנמצאות בשימוש משותף ע"י המחלקות היורשות ממנה.

**המחלקה מכילה את האיברים הבאים:**

1. **SVMLearnLocation**: זהו איבר סטטי של המחלקה המכיל את מיקומו של היישום svm\_learn.exe. יישום זה משמש **ללימוד** המודל עליו מתבססים המסווגים.
2. **SVMClassifyLocation**: זהו איבר סטטי של המחלקה המכיל את מיקומו של היישום svm\_cassify.exe. יישום זה משמש **לסיווג** באמצעות המסווגים.

**המחלקה מכילה את השיטות הבאות:**

1. **GetNetsModelVectorFile**: שיטה זו מקבלת את המיקום של ווקטורי המאפיינים ומחזירה מבנה נתונים המכיל מיפוי משם הרשת לקבצי ווקטורי המאפיינים שלה.
2. **ReadVectorsFile**: שיטה זו קוראת את ווקטורי המאפיינים מקובץ לזיכרון.
3. **ReadAttributesLine**: שיטה זו קוראת ווקטור מאפיינים בודד מקובץ לזיכרון.
4. **NormalizeIfNeeded**: שיטה זו מנרמלת את ווקטור המאפיינים במידה ונדרש. ( בפועל השיטה לא בשימוש ).
5. **RunSVM**: שיטה זו משתמשת במסווג ה-SVM שנבנה, על מנת לסווג קובץ המכיל ווקטורי מאפיינים של רשתות. תוצאות הסיווג נכתבות לקובץ ומנותחות בשלב מאוחר יותר.
6. **CheckTestResult**: שיטה זו בודקת את תוצאות הסיווג, וכותבת את הסיכום שלהן לקובץ.
7. **ComputeTestPrecision**: שיטה זו מחשבת את הדיוק של מסווג מסוים על סמך תוצאות הסיווג.
8. **WriteNetSummaryFile**: שיטה זו מייצרת קובץ המכיל תוצאות מסוכמות של דיוקי כלל המסווגים.
9. **DeleteTrainTest**: שיטה זו מוחקת את קבצי האימון והבדיקה לאחר סיום הבדיקות.

* **המחלקה SimpleHandler**

מחלקה זו אחראית על בניית ובדיקת המסווגים של כלל המודלים. המחלקה יורשת מהמחלקה Handler.

**המחלקה מכילה את השיטות הבאות:**

1. **TrainAndTestClassifiers**: שיטה זו היא השיטה העיקרית של המחלקה. השיטה אחראית על ביצוע כלל הפעולות הנדרשות ליצירת ובדיקת המסווגים.
2. **CreateModelsTrainTestData**: שיטה זו אחראית על ייצור קבצי האימון והבדיקה של כלל המסווגים.
3. **AlreadyDone**: זהי שיטת עזר. מטרתה מניעת יצירה ובדיקה מחודשת של המסווגים. השיטה מסתמכת על השוואת כמות קבצי תוצאות הסיווג לפרמטר שהועבר לשיטה. **יש לשים לב כי פרמטר זה תלוי בכמות המסווגים הנבדקים. לכן במידה וכמות המסווגים משתנה, יש לבצע את ההתאמות הנדרשות לערכו של הפרמטר המועבר**. הואיל ולכל מסווג שנבדק נוצרים שני קבצי תוצאות, ערכו של הפרמטר המועבר צריך להיות שווה לפי שניים מכמות המסווגים הנבדקים.
4. **TrainSVMs**: שיטה זו מאמנת את המסווגים השונים, באמצעות קבצי האימון הרלבנטיים. למטרה זו השיטה משתמשת ביישום svm\_learn.
5. **TestSVMs**: שיטה זו מסווגת את הווקטורים המוכלים בקבצי הבדיקה באמצעות המסווגים שאומנו. תוצאות הסיווג נכתבות לקובץ ומשמשות להערכת הדיוק של המסווגים.
6. **CreateBestClassifiersDir**: שיטה זו מייצרת תיקיה המיועדת להכיל את המסווגים הטובים ביותר מכל סוג ( במידה ועבור כל מסווג נוצרו מספר מסווגים למטרת cross validation ).
7. **WriteNetsSummaryFile**: שיטה זו מייצרת קובץ המכיל את סיכום הנתונים על רמות הדיוק של המסווגים השונים.
8. **WriteNetsBestClassifiersFile**: שיטה זו מייצרת קובץ המכיל את סיכום הנתונים על רמות הדיוק של המסווגים הטובים ביותר.
9. **checkTrainLoss**: שיטה זו משתמשת בקבצי הלוג שנוצרו במהלך אימון המסווגים לחישוב כמות השגיאות שהיו ( train loss ). הפלט נכתב לקובץ מסוג csv.

* **המחלקה RobustnessHandler**

מחלקה זו אחראית לבצע את בדיקת העמידות ( robustness ) של המסווגים השונים שנבנו. אופן הביצוע של פעולה זו דומה מאוד לתהליך יצירת ובדיקת המסווגים עצמם ע"י המחלקה SimpleHandler. ההבדל העיקרי הינו בכך שהסיווג נעשה על תתי קבוצות של המאפיינים

**המחלקה מכילה את השיטות הבאות:**

1. **CheckRobustness**: שיטה זו היא השיטה העיקרית של המחלקה. השיטה אחראית על ביצוע כלל הפעולות הנדרשות לבדיקת העמידות של המסווגים.
2. **AlreadyDone**: זוהי שיטת עזר. מטרתה מניעת חישוב מחדש של עמידות מסווגים במידה והעמידות חושבה כבר. השיטה מסתמכת על השוואת כמות קבצי תוצאות הסיווג לפרמטר שהועבר לשיטה**. יש לשים לב כי פרמטר זה תלוי בכמות המסווגים הנבדקים. לכן במידה וכמות המסווגים השתנתה, יש לעשות את ההתאמות הנדרשות לערכו של הפרמטר המועבר.** הואיל ולכל מסווג שנבדק נוצרים שני קבצי תוצאות, ערכו של הפרמטר המועבר צריך להיות שווה לפי שניים מכמות המסווגים הנבדקים.
3. **CreateTrainTestData**: שיטה זו אחראית על ייצור קבצי האימון והבדיקה של כלל המסווגים עבור תת קבוצה מסוימת של מאפיינים. הביצוע נעשה ע"י קריאה לשיטה CreateTrainTestSet עם הפרמטרים המתאימים.
4. **CreateTrainTestSet**: שיטה זו מייצרת את קבצי האימון והבדיקה של מסווג מסוים עבור תת קבוצה מסוימת של מאפיינים.
5. **CreateAttributesLine**: שיטה זו מאפשרת לייצר מווקטור המאפיינים, תת ווקטור המכיל רק את המאפיינים שנבחרו לשימוש.
6. **TrainSVMs**: שיטה זו מאמנת את המסווגים השונים, באמצעות קבצי האימון הרלבנטיים. למטרה זו השיטה משתמשת ביישום svm\_learn.
7. **TestSVMs**: שיטה זו מסווגת את הווקטורים המוכלים בקבצי הבדיקה באמצעות המסווגים שאומנו. תוצאות הסיווג נכתבות לקובץ ומשמשות להערכת הדיוק של המסווגים.

* **המחלקה NetworkClassifiers**

מחלקה זו אחראית על סיווג הרשתות האמפיריות באמצעות המסווגים שנבנו. לאחר ביצוע הסיווג נכתבים סיכומים שונים של תוצאות הסיווג.

**המחלקה מכילה את השיטות הבאות:**

1. **ClassifyNets**: זוהי השיטה הראשית במחלקה. השיטה אחראית על ביצוע כלל הפעולות הנדרשות לסיווג הרשתות וכתיבת סיכום התוצאות.
2. **CollateClassificationResults**: שיטה זו אחראית לקיבוץ התוצאות של כלל הסיווגים של הרשתות, על מנת שיהיה ניתן לכתוב אותם באופן סיכומי.
3. **WriteResultSummary**: שיטה זו משמשת לכתיבת תוצאות הסיווג הסיכומיות עבור כל מסווג.
4. **WriteSummary**: שיטה זו משמשת לכתיבת תוצאות הסיווג בצורה של מטריצת הצבעות. במטריצה זו ניתן לראות עבור כל **מודל,** כמה מודלים הוא "ניצח", ומה אחוז הניצחונות שלו מול כל מודל מכלל המסווגים בהם הוא השתתף.

* **המחלקה Program**

המחלקה Program אחראית על הפעולה שייבצע היישום.

**המחלקה מכילה את השיטות הבאות:**

1. **Main**: קריאת הפרמטרים איתם הורץ היישום והחלטה על ביצוע הפעולה המתאימה.

### הרצת מסווג הרשתות

על מנת להשתמש ביישום ניתן להריץ אותו באופן הבא:

**NetworksClassifiers** –a <A,B,R,N,L> -sm <model vectors dir>

-sn <networks vectors dir> -d < destination dir> -v <vectors amount>

-t <train examples amount> -c < cross Validation amount> -rs <robustness samples>

-ra <robustness attributes amount> -n <0,1>

**יש לשים לב לנקודות הבאות:**

* ההרצה משתמשת בזוגות פרמטרים: הראשון מצביע על סוג הפרמטר והשני על הערך שלו.
* אין חשיבות לסדר הזוגות של הפרמטרים.
* יש להוסיף רווח בין סוג הפרמטר לבין הערך שלו.
* אין חובה לרשום את כל הפרמטרים. במידה ופרמטר לא נרשם יבוצע שימוש בערך ברירת המחדל שלו.
* **חשוב לזכור:**  יישום זה כבד מאוד מבחינת משאבים. הן מבחינת זמן הריצה והן מבחינת המקום הפנוי הנדרש בדיסק הקשיח. רצוי מאוד לנטר מפעם לפעם את פעילות היישום ולוודא שהוא לא נתקל בבעיות.
* היישום נכתב כך שתהיה לו יכולת התאוששות מסויימת. כלומר, במידה והיישום "נפל" מסיבה כלשהי, אזי בפעם הבאה שהוא יורץ הוא יינסה להמשיך את פעולתו מנקודת הנפילה.
* **חשוב לזכור:**  ישנן תלויות בין הפעולות השונות שיישום זה מבצע. הפעולה הראשונה **חייבת** להיות בניית המסווגים. אח"כ רצוי להריץ את בדיקת העמידות ורק לבסוף את סיווג הרשתות.

**משמעות הפרמטרים היא:**

1. **-a**: סוג הפעולה שיש לבצע, כאשר האפשרויות הן:
   1. **B**: בניית המסווגים ובדיקת האיכות שלהם.
   2. **R**: בדיקת עמידות המסווגים.
   3. **N**: סיווג הרשתות האמפיריות.
   4. **L**: בדיקת אובדן האיכות (train loss) במהלך בניית המסווגים.
   5. **A**: בצע את כל הפעולות.
2. **–sm:** מיקום התיקייה המכילה את ווקטורי המאפיינים של המודלים.
3. **–sn**: מיקום התיקייה המכילה את ווקטורי המאפיינים של הרשתות האמפיריות.
4. **–d**: מיקום התיקייה אליה ייכתבו התוצאות.
5. **–v**: כמות ווקטורי המאפיינים שיש בכל קובץ (היישום מניח כי כמות ווקטורי המאפיינים זהה עבור כל המודלים).
6. **–t**: מספר ווקטורי מאפיינים בהם יש להשתמש עבור אימון המסווגים (הווקטורים הנותרים ישמשו לבדיקת המסווג שנבנה ).
7. **–c**: כמות הבדיקות המוצלבות (cross validation) שיש לעשות. הערכים האפשריים הם מספרים שלמים בין 1 ל v/t. ערכים גדולים יותר ייגרמו לכך שהבדיקות תחזורנה על עצמן.
8. **–rs**: כמות בדיקות העמידות שיש לבצע עבור כל מסווג. בכל בדיקת עמידות נבדקת תת קבוצה שונה של מאפיינים הנבחרת באקראי.
9. **–ra**: גודל תת קבוצת המאפיינים הנבדקת בכל בדיקת עמידות.
10. **–n**: האם לנרמל את ווקטורי המאפיינים. ( 0-לא, 1- כן ).

**דוגמא:**

הפקודה הבאה תגרום ליישום לבצע את כל הפעולות עם הנתונים הבאים:

* ווקטורי המאפיינים נמצאים בספריה C:\ModelVectors.
* התוצאות יישמרו בספריה C:\Results.
* כמות הווקטורים בכל קובץ הינה 500, מהם 100 ישמשו לאימון והשאר לבדיקה.
* הבדיקה תבוצע 5 פעמים על קבוצות שונות.
* למסווגים שנבנו יבוצעו 30 בדיקות עמידות, כל אחת על 500 מאפיינים.
* בסיום התהליך יסווגו רשתות האמת הנמצאות בתיקיה C:\NetworkVectors באמצעות המודלים שנבנו ותוצאות הסיווג יירשמו לקובץ.

**NetworksClassifiers** –a A –sm c:\modelVectors –sn c:\networkVectors

–d c:\Results –v 500 –t 100 –c 5 –rs 30 –ra 500 -n 0

### תיאור התוצאות

**בחירה לבנות את המסווגים תייצר תחת תיקיית היעד את עץ התיקיות הבא:**

* **TrainTest**: תיקיה זו אמורה להיות ריקה בסוף ריצת התוכנית. במהלך התוכנית נכתבים לתיקיה זו קבצים זמניים המשמשים ללימוד ובדיקת המסווגים.
* **SVMModels**: תיקיה זו מכילה את המודלים של המסווגים שנוצרו באמצעות ה-svm\_learn. בנוסף מכילה התיקייה את קבצי הלוג של תהליך ה-svm\_learn. ( אם כי אין בהם מידע מעניין ).
* **BestSVMModels**: תיקיה זו מכילה עבור כל זוג מודלים, את המסווג הטוב ביותר ביניהם (כזכור, ניתן ליצור מספר מסווגים עבור כל זוג מודלים לשם בדיקה מוצלבת ( cross validation ).
* **SVMTestResult**: תיקיה זו מכילה את תוצאות הבדיקה של המסווגים שנבנו על קבצי הבדיקה. בנוסף, מכילה התיקייה שני קבצי csv הראשון בשם: Classifiers.Summary.Result.csv המכיל את הדיוק (precision) של כל אחד מהמסווגים והשני בשם: BestClassifiers.Summary.Result.csv מכיל את תוצאות הדיוק של המסווגים הטובים ביותר.

**בחירה לבצע את בדיקת העמידות** תייצר תחת תיקיית היעד תיקיה בשם Robustness. תחת תיקיה זו ייווצר עץ תיקיות הדומה לעץ התיקיות שנוצר עבור בניית המסווגים. ההבדל בין שני עצי התיקיות הינו בכך שעבור בדיקת העמידות לא ייווצרו תיקיות עבור המסווגים הטובים ביותר הואיל ובמקרה זה לא מתבצעת בדיקה מוצלבת ( cross validation ) ולכן דבר שכזה הינו חסר משמעות. כמו-כן, ייווצרו בתיקיה קבצים המכילים את רשימת המאפיינים שהייתה בשימוש עבור כל אחת מבדיקות העמידות.

**בחירה לסווג את הרשתות תייצר תחת תיקיית היעד את עץ התיקיות הבא:**

* **TrueNetsClassificationResult**: תיקיה זו מכילה את תוצאות הסיווג של הרשתות האמפיריות באמצעות כלל המסווגים. כמו-כן מכילה התיקייה שני קבצי csv: הראשון עם תוצאות מפורטות על תוצאות הסיווג לפי כל מודל והשני עם ריכוז כמות הניצחונות של כל מודל.
* **TrueNetsBestClassificationResult**: תיקיה זו מכילה את תוצאות הסיווג של הרשתות האמפיריות באמצעות המסווגים הטובים ביותר. כמו-כן מכילה התיקייה שני קבצי csv: הראשון עם תוצאות מפורטות על תוצאות הסיווג לפי כל מודל והשני עם ריכוז כמות הניצחונות של כל מודל.

## ב.3. בודק המובהקות ( Significance Finder )

### מבנה

המבנה של בודק המובהקות מורכב מן המחלקות הבאות:

* **המחלקה ClassifierPercision**

מחלקה זו מחזיקה נתונים על הדיוק של מסווג מסוים. על מנת להבין יותר את משמעות האיברים של המחלקה , יש לזכור כי כל מסווג מבחין בין שני מודלים. לצורך אבחנה בין המודלים נקרא להם המודל ה"שמאלי" והמודל ה"ימני". הסיווג של דוגמת בדיקה ע"י המסווג ניתן ע"י ערך מספרי. ערך חיובי משמעו כי המסווג קובע שדוגמת הבדיקה שייכת למודל ה"שמאלי", ואילו ערך שלילי משמעו כי המסווג קובע שדוגמת הבדיקה שייכת למודל ה"ימני".

**המחלקה מכילה את האיברים הבאים:**

1. **Precision**: הדיוק של המסווג על כלל קבוצת הבדיקה.
2. **PositivePrecision**: הדיוק של המסווג על הווקטורים בקבוצת הבדיקה שהיו שייכים למודל ה"שמאלי"
3. **NegativePrecision**: הדיוק של המסווג על הווקטורים בקבוצת הבדיקה שהיו שייכים למודל ה"ימני"
4. **TruePositive**: אחוז דוגמאות הבדיקה שהיו שייכות למודל ה"שמאלי" וסווגו נכונה.
5. **FalsePositive**: אחוז דוגמאות הבדיקה שהיו שייכות למודל ה"שמאלי" וסווגו באופן שגוי.
6. **TrueNegative**: אחוז דוגמאות הבדיקה שהיו שייכות למודל ה"ימני" וסווגו נכונה.
7. **FalseNegative**: אחוז דוגמאות הבדיקה שהיו שייכות למודל ה"ימני" וסווגו באופן שגוי.

**המחלקה מכילה את השיטות הבאות:**

1. **בנאי**: הבנאי מקבל את ערכי ה-Precision,PositivePrecision ו-NegativePrecision ומחשב את שאר איברי המחלקה.
2. שיטות לקבלת ערכי האיברים של המחלקה.

* **המחלקה SignificanceFinder**

זוהי המחלקה העיקרית של היישום. המחלקה מקבלת את קבצי המידע של תוצאות הסיווג, העמידות ואיכות המסווגים, ומחשבת את המובהקות של התוצאות.

**המחלקה מורכבת מהשיטות הבאות:**

1. **Run**: שיטה זו אחראית לבצע את כלל הפעולות הנדרשות לחישוב המובהקות. השיטה מקבלת את מיקום קבצי הנתונים, קוראת אותם, מסננת את נתוני המסווגים שלא היו בעלי ערכי עמידות גבוהים כנדרש, ומחשבת את המובהקות של הסיווג עבור כל אחד מהמסווגים.
2. **ReadResultFile**: שיטה זו קוראת את קובץ תוצאות הסיווג לזיכרון.
3. **ReadPrecisions**: שיטה זו קוראת את הקובץ המכיל את הדיוק של המסווגים השונים לזיכרון.
4. **FilterClassifications**: שיטה זו מנפה מכלל תוצאות הסיווג את התוצאות של המסווגים שלא עמדו בתנאים הבאים:
   1. ערך העמידות שלהם לא היה מספיק גבוה (מועבר כפרמטר ע"י המשתמש).
   2. תוצאת הסיווג הייתה שונה מתוצאת הסיווג של תתי המסווגים.
   3. היה רוב **ברור** (מועבר כפרמטר ע"י המשתמש ) לתוצאת הסיווג ( סעיף זה רלבנטי רק למקרה של טיפול במסווגים מרובים הנוצרים כתוצאה מביצוע cross validation ).
5. **AssignNetIds**: שיטה זו מבצעת מיפוי של שמות הרשתות שסווגו למספרים. מיפוי זה נדרש על מנת שיהיה ניתן לבחור תתי קבוצות של רשתות בצורה יותר נוחה.
6. **CalculateSignificance**: שיטה זו מבצעת בפועל את חישוב מובהקות הסיווג של כלל המסווגים.
7. **CalcClassifierSignificance**: שיטה זו מחשבת את המובהקות של מסווג מסוים, עבור כמות שגיאות מרבית מסוימת.
8. **GetAvgPrecisions**: שיטה זו מחשבת את הדיוק הממוצע של המסווגים. חישוב זה נדרש במקרה בו חישוב המובהקות מבוצע על מספר מסווגים שנוצרו כתוצאה מיצירת מספר מסווגים השייכים לאותו סוג ועל אותה רשת. מסווגים אלו נוצרים עקב ביצוע בדיקה מוצלבת ( cross validation ).
9. **CalcSubsetsSignificance**: שיטה זו מחשבת את ערך המובהקות, כאשר מניחים כי חלק מתוצאות הסיווג היו שגויות. החישוב נעשה ע"י מעבר על כל תתי הקבוצות של סיווגים בגדלים הרלבנטיים ובמידה ומניחים כי הסיווג היה מוטעה, משתמשים בדיוק של המקרה המשלים.
10. **GetGroupSignificance**: שיטה זו מחשבת את ציון המובהקות של תת קבוצה מסוימת של סיווגים.
11. **GetGroup**: שיטה זו מחזירה את ערכי הדיוק, של קבוצת סיווגים מסוימת, על מנת שיהיה ניתן לחשב את ציון המובהקות שלה.
12. **Advance**: שיטה זו עוברת לתת הקבוצה הבאה של סיווגים. השיטה מחזירה false במקרה וסיימנו לעבור על כל תת הקבוצות.

* **המחלקה Program**

המחלקה Program אחראית על קריאת הפרמטרים איתם הורץ היישום, ניתוחם וקריאה לשיטה Run של המחלקה SignificanceFinder.

**המחלקה מכילה את השיטות הבאות:**

1. **Main**: קריאת הפרמטרים איתם הורץ היישום והחלטה על ביצוע הפעולה המתאימה.

### הרצת בודק המובהקות

על מנת להשתמש ביישום ניתן להריץ אותו באופן הבא:

**SignificanceFinder** –cr <filename> -cp <filename> -r <filename> -o <filename>

-mm <max mistakes> -mr <min robustness>

**יש לשים לב לנקודות הבאות:**

* ההרצה משתמשת בזוגות של פרמטרים: הראשון מצביע על סוג הפרמטר והשני על הערך שלו.
* אין חשיבות לסדר הזוגות של הפרמטרים.
* יש להוסיף רווח בין סוג הפרמטר לבין הערך שלו.
* **חשוב לזכור**: ככל שכמות השגיאות המרבית גדלה, כך זמן הריצה של היישום מתארך. זמן הריצה גדל גם כתלות בכמות הרשתות עליהן מתבצעת הבדיקה הואיל ויש צורך לעבור על יותר תת קבוצות.

**משמעות הפרמטרים היא:**

1. **-cr**: מיקום הקובץ המכיל את תוצאות סיווג הרשתות.
2. **–cp:** מיקום הקובץ המכיל את הדיוק של המסווגים.
3. **–r**: מיקום הקובץ המכיל את תוצאות בדיקת העמידות של המסווגים.
4. **–o**: שם קובץ הפלט אליו ייכתבו התוצאות.
5. **–mm**: כמות שגיאות הסיווג המירבית עבורה רוצים לבדוק את רמת המובהקות.
6. **–mr**: ערך העמידות המינימאלי לפיו יש לסנן את תוצאות הסיווג.

**דוגמא:**

הפקודה הבאה תגרום להרצת היישום כך ש:

* קובץ תוצאות הסיווג יילקח מ- C:\ClassificationResults.
* קובץ דיוק המאפיינים יילקח מ-C:\ClassifierPercisions.
* קובץ תוצאות בדיקת העמידות יילקח מ-C:\RobustnessResults.
* התוצאות יכתבו ל-C:\SignificanceResults.
* תיבדק המובהקות של לכל היותר 3 שגיאות סיווג
* ערך העמידות המינימאלי יהיה 90%.

**SignificanceFinder** –cr c:\ClassificationResults –cp c:\ClassifiersPercisions –r C:\RobustnessResult –o C:\SignificanceResults –mm 3 –mr 0.9

### תיאור התוצאות

הפלט של היישום הינו קובץ המכיל את מובהקות הסיווג עבור המסווגים השונים. קובצי התוצאות הינו קובץ csv המורכב ממספר עמודות:

1. שם המסווג.
2. מובהקות הסיווג.
3. כמות שגיאות סיווג מרבית.
4. כמות נתונים עליהם בוצעה הבדיקה – כלומר כמות הנתונים שעברו את הסינון בהתאם להגבלות על עמידות המסווגים.
5. כמות הנתונים מתוך אלו עליהם בוצעה הבדיקה בהם ניצח המודל המנצח
6. שם המודל המנצח.

**דוגמא למבנה קובץ התוצאות**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Significance | MaxMistakes | ExamplesAmount | Wins | Winner |
| DRG/SPA | 0.613166186 | 0 | 35 | 35 | SPA |
| SPA/DPA | 0.219448389 | 0 | 28 | 28 | SPA |
| SPA/DC | 1 | 0 | 35 | 35 | SPA |
| SPA/DSW | 1 | 0 | 34 | 34 | SPA |
| SPA/IDPA | 0.001112413 | 0 | 35 | 35 | SPA |
| DRG/SPA | 0.920193922 | 1 | 35 | 35 | SPA |
| SPA/DPA | 0.589905752 | 1 | 28 | 28 | SPA |
| SPA/DC | 1 | 1 | 35 | 35 | SPA |
| SPA/DSW | 1 | 1 | 34 | 34 | SPA |
| SPA/IDPA | 0.009910516 | 1 | 35 | 35 | SPA |

# מקורות

1. Adamic, L. A. and Huberman, B. A., (2000). *Power-law distri-bution of the world wide web,* Science 287, 2115.
2. Albert, R., Jeong, H., and Barabasi, A.-L., (1999). *Diameter of the world-wide web*, Nature 401, 130–131.
3. Albert, R. And Barabasi, A.-L., (1999). *Emergence of scaling in random networks*, Science 286, 509–512.
4. Albert, R., Jeong, H., and Barabasi, A.-L., (2000). *Attack and error tolerance of complex networks*, Nature 406, 378–382.
5. Albert, R., & Barabasi, A. L. (2002). *Statistical Mechanics of Complex Networks.* Review of Modern Physics, 74, 47-97.
6. Aviv, R., Erlich, Z., Ravid, G., & Geva, A. (2003). *Network Analysis of Knowledge Construction in Asynchronous Learning Networks. Journal of Asynchronous Networks*, 7(3), 1-23.
7. Aviv, R., Erlich, Z., & Ravid, G. (2007a). *Analysis of Reciprocity and Transitivity in Online Collaboration Networks. Connections*.
8. Aviv, R., Erlich, Z., & Ravid, G. (2007b). *Randomness and Clustering of Responses in Online Learning Networks*. Paper presented at the IASTED International Conference on Communication, Internet and Information Technology.
9. Bekessey, A., Bekessey, P., and Komlos J., (1972). *Asymptotic enumeration of*  *regular matrices*, Stud. Sci. Math. Hugar. 7, 343-353.
10. Bender, E. A., and Canfield, E. R., (1978). *The asymptotic number of labeled graphs with given degree sequences.* Journal of Combinatorical Theory A 24, 296-307.
11. Bianconi, G. and Barabasi, A.-L., (2001). *Competition and multiscaling in evolving networks*, Europhys. Lett. 54, 436–442.
12. Bollob´as, B. (1980). *A probabilistic proof of an asymptotic formula for the number of labeled regular graphs,* European Journal on Combinatorics 1, 311-316.
13. Bollob´as, B. and Riordan, O. (2002). *The diameter of a scale-free random graph*, Preprint, Department of Mathematical Sciences, University of Memphis.
14. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J., (2000). *Graph structure in the web*, Computer Networks 33, 309–320.
15. Buts, C. T., (2008), Social Network Analysis, A Methodological Introduction, Asian Journal of Soc. Psych, 11, 13-41
16. Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J., and Strogatz, S. H., (2001). *Are randomly grown graphs really random?*, Phys. Rev. E 64, 041902.
17. Chung, F. and Lu, L., (2002). *The average distances in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA 99, 15879–15882.
18. Dorogovtsev, S. N. and Mendes, J. F. F., (2000). *Scaling behaviour of developing and decaying networks*, Europhys. Lett. 52, 33–39.
19. Dorogovtsev, S. N. and Mendes, J. F. F., (2001). *Effect of the accelerating growth of communications networks on their structure*, Phys. Rev. E 63, 025101.
20. Dorogovtsev, S. N., Mendes, J. F. F., and Samukhin, A. N., (2001). *Anomalous percolation properties of growing networks*, Phys. Rev. E 64, 066110.
21. Dunne, J. A., Williams, R. J., and Martinez, N. D., (2002). *Food-web structure and network theory: The role of connectance and size*, Proc. Natl. Acad. Sci. USA 99, 12917–12922.
22. Erdos, P., & R'enyi, A. (1960). *On the Evolution of Random Graphs*. Publ. Math. Inst. Hung. Acad. Sci., 5, 17–61.
23. Everitt, B., (1974). *Cluster Analysis*, John Wiley, New York.
24. Goh, K.-I., Khang, B., & Kim, D. (2001). *Universal behavior of load distribution in scale free networks.* *Phys. Rev. Lett., 87*(27).
25. Goh, K.-I., Oh, E., Jeong, H., Kahng, B., and Kim, D., (2002). *Classification of scale-free networks*, Proc. Natl. Acad. Sci. USA 99, 12583–12588.
26. Jeong, H., Mason, S., Barabasi, A.-L., and Oltvai, Z. N. (2001). *Lethality and centrality in protein networks*, Nature 411, 41–42.
27. Jeong, H., N´eda, Z., and Barabasi, A.-L., (2003). *Measuring preferential attachment in evolving networks*, Europhys. Lett. 61, 567–572.
28. Joachims, T. (1999). *Making large-scale SVM Learning Practical*. In B. Schlopf, C. Burges & A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Machines*: MIT-Press.
29. Kleinberg, J. M., Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A., (1999). *The Web as a graph: Measurements, models and methods*, in Proceedings of the International Conference on Combinatorics and Computing, no. 1627 in Lecture Notes in Computer Science, pp. 1–18, Springer, Berlin.
30. Krapivsky, P. L. and Redner, S., (2001). *Organization of growing random networks*, Phys. Rev. E 63, 066123.
31. Krapivsky, P. L., Rodgers, G. J., & Redner, S., (2001). *Degree distributions of growing networks*. *Phys. Rev. Lett., 86*(23), 5401-5404.
32. Krapivsky, P. L. and Redner, S., (2002). *A statistical physics perspective on Web growth*, Computer Networks 39,261–276.
33. Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. S., and Upfal, E., (2000). *Stochastic models for the Web graph,* in Proceedings of the 42st Annual IEEE Symposium on the Foundations of Computer Science, pp. 57–65, Institute of Electrical and Electronics Engineers, New York.
34. Maslov, S. and Sneppen, K., (2002). *Specificity and stability in topology of protein networks*, Science 296, 910–913.
35. Middendorf, M., Ziv, E., Adams, C., Hom, J., Koytcheff, R., Levovitz, C., et al. (2004). *Discriminative Topological Features Reveal Biological Network Mechanisms*. *BMC Bioinformatics, 5*, 181.
36. Milgram, S. (1967). *The small world problem*, Psychology To-day 2, 60–67.
37. Newman, M. E. J., (2001). *Clustering and preferential attachment in growing networks*, Phys. Rev. E 64, 025102.
38. Newman, M. E. J., (2002). *Assortative mixing in networks*, Phys. Rev. Lett. 89, 208701.
39. Newman, M. E. J., Forrest, S., and Balthrop, J., (2002). *Email networks and the spread of computer viruses*, Phys. Rev. E 66, 035101.
40. Newman, M. E. J. (2003). *The Structure and Function of Complex Networks*. *SIAM Review, 45*(2), 167-256.
41. Pastor-Satorras, R., V´azquez, A., and Vespignani, A., (2001). *Dynamical and correlation properties of the Internet*, Phys. Rev. Lett. 87, 258701.
42. Price, D. J. de S., (1965). *Networks of scientific papers*, Science 149, 510–515.
43. Price, D. J. de S., (1976). *A general theory of bibliometric and other cumulative advantage processes*, J. Amer. Soc. Inform. Sci. 27, 292–306.
44. Rapoport, A., (1957). *Contribution to the theory of random and biased nets*, Bulletin of Mathematical Biophysics 19, 257–277.
45. Scott, J., (2000). *Social Network Analysis: A Handbook*, Sage Publications, London, 2nd Ed.
46. Shen-Orr, S., Milo, R., Mangan, S., & Alon, U. (2002). *Network motifs in the transcriptional regulation network of Escherichia coli*, Nature Genetics, **31**, 64-68.
47. Simon, H. A., (1955). *On a class of skew distribution functions*, Biometrika 42, 425–440.
48. Strogatz, S. H., (2001). Exploring complex networks, Nature 410, 268–276.
49. Tadi´c, B., (2001). *Dynamics of directed graphs: The World-Wide Web*, Physica A 293, 273–284.
50. Vapnik, V., (1995). *The Nature of Statistical Learning Theory*. NY: Springer-Verlag.
51. Vazquez, A. (2002). *Knowing a network by walking on it: Emergence of scaling.* *arXiv.org: cond-mat/0006132*.
52. Watts, D. & Strogatz, S., (1998). *Collective dynamics of small-world networks*, Nature **363**, 202–204.
53. Wasserman, S & Faust, K., (1994). *Social Network Analysis*, Cambridge University Press, Cambridge.

|  |
| --- |
| The Open University of Israel  Department of Mathematics and Computer Science  Revealing Mechanisms in Online Learning Networks |
|  |
| Thesis submitted as partial fulfillment of the requirements  towards an M.Sc. degree in Computer Science  The Open University of Israel  Computer Science Division |
|  |
| By  **Moshe Mazuz** |
|  |

Prepared under the supervision of Prof. Reuven Aviv

|  |
| --- |
|  |

October 2009

**Content**

[Abstract 1](#_Toc218416208)

[1 Introduction 2](#_Toc218416209)

[2 Complex networks research background 5](#_Toc218416210)

[2.1 General definitions 5](#_Toc218416211)

[2.2 Properties of networks 6](#_Toc218416212)

[2.2.1 Small World 6](#_Toc218416213)

[2.2.2 Transitivity or clustering 6](#_Toc218416214)

[2.2.3 Degree Distribution 8](#_Toc218416215)

[2.2.4 Network resilience 9](#_Toc218416216)

[2.2.5 Mixing patterns 10](#_Toc218416217)

[2.2.6 Degree correlations 11](#_Toc218416218)

[2.2.7 Community structure 11](#_Toc218416219)

[2.2.8 Betweens Centrality 12](#_Toc218416220)

[2.3 The Random graph model 12](#_Toc218416221)

[2.3.1 Model definition 13](#_Toc218416222)

[2.3.2 The effect of connection probabily on the graph structure 13](#_Toc218416223)

[2.3.3 The Random graph model Vs Networks in real world 14](#_Toc218416224)

[2.3.4 Generalized random graphs - Configuration model 15](#_Toc218416225)

[2.4 Small world model 15](#_Toc218416226)

[2.4.1 Model properties 16](#_Toc218416227)

[2.5 Preferential attachment models 16](#_Toc218416228)

[2.5.1 The Price model 17](#_Toc218416229)

[2.5.2 The Barabasi & Albert model 19](#_Toc218416230)

[2.6 Vertex copying and change models 23](#_Toc218416231)

[2.6.1 Attaching to edges mechanism 23](#_Toc218416232)

[2.6.2 Vertex copying mechanism 24](#_Toc218416233)

[2.6.3 Walking on a network 25](#_Toc218416234)

[3 Research goal 26](#_Toc218416235)

[4 The database 27](#_Toc218416236)

[5 Network generation models 28](#_Toc218416237)

[5.1 Connectivity testing models – Binary models 28](#_Toc218416241)

[5.1.1 Directed randon graph model 28](#_Toc218416242)

[5.1.2 Static preferentail attachment model 29](#_Toc218416243)

[5.1.3 Dynamic preferentail attachment model 30](#_Toc218416244)

[5.1.4 Improved Dynamic preferentail attachment model 32](#_Toc218416245)

[5.1.5 Dynamic copying model 33](#_Toc218416246)

[5.1.6 Small world model ( direced version ) 34](#_Toc218416247)

[5.2 Weighted models 35](#_Toc218416248)

[5.2.1 Weighted directed random graph 35](#_Toc218416249)

[5.2.2 Weighted static preferentail attachment model 36](#_Toc218416250)

[5.2.3 Weighted dynamic preferentail attachment model 36](#_Toc218416251)

[5.2.4 Weighted improved dynamic preferentail attachment model 36](#_Toc218416252)

[5.2.5 Weighted static dynamic preferentail attachment model 37](#_Toc218416253)

[6 Methodology 38](#_Toc218416254)

[6.1 Classifiers creation 38](#_Toc218416255)

[6.2 Classifiers robustness check 39](#_Toc218416256)

[6.3 Choosing classifiers 39](#_Toc218416257)

[6.4 Identifying the winning model 39](#_Toc218416258)

[6.5 Results confidence level 40](#_Toc218416259)

[7 Results 42](#_Toc218416260)

[7.1 Connectivity results 42](#_Toc218416261)

[7.1.1 Classifying 42](#_Toc218416262)

[7.1.2 Robustness check 46](#_Toc218416263)

[7.1.3 Identifying the winning model 48](#_Toc218416264)

[7.1.4 Results confidence level 50](#_Toc218416265)

[7.2 Weighted Case 52](#_Toc218416266)

[7.2.1 Classifying 52](#_Toc218416267)

[7.2.2 Identifying the winning model 55](#_Toc218416268)

[7.2.3 Robustness check 57](#_Toc218416269)

[7.2.4 Results confidence level 59](#_Toc218416270)

[8 Discussion 61](#_Toc218416271)

[9 Summary and directions for future research 64](#_Toc218416272)

[Appendix A: Network representation in multi dimensional space 66](#_Toc218416273)

[Description 66](#_Toc218416274)

[Advantages 68](#_Toc218416275)

[Disadvantages 68](#_Toc218416276)

[Appendix B: Software description 69](#_Toc218416277)

[B.1: Attributes Creator 69](#_Toc218416277)

[Structure 69](#_Toc218416277)

[Running the AttributesCreator 77](#_Toc218416277)

[Description of the networks files 78](#_Toc218416277)

[Results description 78](#_Toc218416277)

[B.2: Networks Classifier 79](#_Toc218416277)

[Structure 79](#_Toc218416277)

[Running the NetworksClassifier 82](#_Toc218416277)

[Results description 83](#_Toc218416277)

[B.3: Significance Finder 85](#_Toc218416277)

[Structure 85](#_Toc218416277)

[Running the Significance Finder 87](#_Toc218416277)

[Results description 88](#_Toc218416277)

[References 89](#_Toc218416278)

Abstract

Online Learning Networks are used as knowledge building tool. These networks are created by interaction between people belong to a learning community connected through the Internet. The goal of this research is to identify response mechanisms in these networks. We ask whether actors choose their response partners at random or whether certain special mechanisms are at work. In that case, we would like to discover what mechanism is most descriptive of the networks. While previous studies checked a few selective attributes of the networks, in this research, we capture their rich complex feature space by mapping them into a high-dimensional feature space. A Multi-way Support Vector Machine algorithm is used to classify 35 observed response networks of online learners. The networks are classified twice: first, they are classified into a set of six representative stochastic network generation models, by using the connectivity information only. Second, they are classified into a set of five representative stochastic network generation models, by using the connection weights information. The results show that in the connectivity check, all the response networks were classified into a *preferential response* model in which actors tend to respond to partners who are a-priori equipped with response attraction power. The results were similar when we checked the connection weights, though this time small part of the networks where classified to other models. We provide a possible explanation for this behavior, based on the nature and goal of the online learning networks, and discuss ways in which the study can continue.

Parts of this work were presented at the following conferences:

1. **Chais 2009**: Learning in the Technological Era, Chais Conference on Instructional Technologies Research, Open University of Israel, February, 2009.
2. **WBE 2009** - The Eighth IASTED International Conference on Web-based Education, Phuket, Thailand, March 2009.