



**EDUCACIÓN CON
RESPONSABILIDAD
SOCIAL**

**USO DE LA COBERTURA PARA
DETERMINAR EL TAMAÑO DE UNA SUBRED**

TESIS

**QUE PARA OBTENER EL GRADO DE
LICENCIADO EN MATEMÁTICAS**

PRESENTA

**YAIR ANTONIO CASTILLO CASTILLO
A LA FACULTAD DE CIENCIAS
DE LA UNIVERSIDAD DE COLIMA**

ASESOR:

**DR. CARLOS MOISÉS HERNÁNDEZ SUÁREZ
JUNIO 2019**

Índice general

Índice general	II
Índice de figuras	IV
Resumen	IV
Abstract	V
1. Introducción	1
2. Preliminares	4
2.1. Redes de Watts-Strogatz	4
2.2. Singletons	6
2.3. Cobertura	7
2.4. Diseños de Link-Tracing y muestreo de bola de nieve (SnowBall Sampling)	11
3. Técnica de muestreo de redes utilizando generaciones	13
3.1. Hipótesis de la red de la población y de la subred del grupo difícil de encontrar.	13
3.2. Descripción de la técnica	14
3.2.1. Simulaciones	17
3.2.2. Observaciones	19
3.3. Mejora de la técnica agregando el término Lag	19
3.3.1. Simulaciones	19
3.3.2. Observaciones	22
3.4. ¿Qué pasa con Lag_∞ ?	22
3.4.1. Simulaciones	22
3.4.2. Observaciones	24
4. Técnica de muestreo utilizando caminata aleatoria	25
4.1. Descripción de la técnica	25
4.1.1. Simulaciones	29
4.1.2. Observaciones	31

4.2. Mejora de la técnica utilizando la muestra como el número de nodos muestreado	31
4.2.1. Simulaciones	31
4.2.2. Observaciones	33
5. Conclusiones	34
Bibliografía	37

Índice de figuras

2.1.	Ejemplo de red de Watts-Strogatz.	5
2.2.	Efecto de p en la topología de la red	6
2.3.	Ejemplo de singletons	6
2.4.	Ejemplo de cobertura y muestra del ejemplo de la cobertura	8
3.1.	(a) Red H , (b) Subred G , (c) G enumerada, (d) Nodo que conocemos, (e) Nodos que conocen el primero, (f) Nodos nuevos	15
3.2.	(g) Nodos que conocen los nodos nuevos, (h) Nodos nuevos, (i) Nodos que conocen los nodos nuevos (j) Nodos nuevos, (k) Nodos que conocen los nodos nuevos	16
3.3.	Grafica de real contra estimada	17
3.4.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$	18
3.5.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$ utilizando $Lag = 1$	20
3.6.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$ utilizando $Lag = 10$	21
3.7.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$	23
4.1.	(a) Red H, (b) Subred G con el nodo que conocemos, (c) Nodos que se conectan con el primero, (d) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (e) Nodos que se conectan al nodo elegido al azar, (f) Nodos nuevos y nodo con flecha roja es el que se elige al azar	27
4.2.	(g) Nodos que se conectan al nodo elegido al azar, (h) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (i) Nodos que conocen el nodo elegido al azar, (j) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (k) Nodos que conocen el nodo elegido al azar, (l) Nodos nuevos y nodo con flecha roja es el que se elige al azar	28
4.3.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 100 nodos con coeficiente $k = 3$	30
4.4.	Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 100 nodos con coeficiente $k = 3$	32

Resumen

En esta tesis se propone utilizar un esquema de muestreo poco conocido, que se aplica generalmente cuando una población tiene dos tipos de individuos, A y B, y se desea contabilizar cuantos individuos hay del tipo A. Dado que no es posible conocer el tipo de un individuo hasta que se entrevista, obtener el conteo del número de individuos de un tipo es una tarea que requiere censar toda la población. Se propone aprovechar la estructura de conexión que hay entre las personas de tipo A, utilizando una expresión matemática conocida como la cobertura de una muestra. Los métodos que se pueden emplear son “SnowBall sampling” y “Random walk sampling” en el que, bajo ciertos supuestos, los individuos de tipo A forman una red. Así, un individuo nos puede dirigir a otros individuos del mismo tipo ahorrando tiempo en teoría. Sin embargo, las poblaciones pueden ser muy grandes y no existe una metodología para evaluar cuanto se ha avanzado en la búsqueda de todos los individuos de un grupo, es decir, si el muestreo se suspende, es imposible hoy en día proponer una forma de estimar el porcentaje de individuos que se han identificado. En este trabajo se propone un criterio para desarrollar la muestra que se prueba bajo diferentes supuestos. Se hace uso de modelos de conexión entre individuos del tipo “Small-World” y de la teoría de la cobertura de una muestra para decidir si se detiene o no el muestreo y evitar navegar la red a ciegas.

Abstract

In this thesis we propose to use a little-known sampling scheme, which is generally applied when a population has two types of individuals, A and B, and we want to count how many individuals there are of type A. Since it's not possible to know the type of an individual until is interviewed, obtaining the number of individuals of a type is a task that requires the census of the entire population. We propose to take advantage of the connection structure between people of type A, using a mathematical expression known as the coverage of a sample. The methods that can be used are "SnowBall sampling" and "Random-walk sampling" in which, under certain assumptions, type A individuals form a network. Thus, an individual can direct us to other individuals of the same type saving time in theory. However, populations can be very large and there is no methodology to assess how much progress has been made in the search for all individuals in a group. That is, if sampling is suspended, it is impossible today to propose a way to estimate the percentage of individuals that have been identified. We proposed a criterion to develop the sample that is tested under different assumptions. Use is made of connection models between individuals of the "Small-Worlds" type and of the theory of the coverage of a sample. In addition to decide whether to stop or not the sample and avoid surfing the network blindly.

Capítulo 1

Introducción

En muchas ocasiones es deseable tener un catálogo de individuos que pertenecen a un grupo de personas en particular, por ejemplo: individuos que consumen drogas, individuos que han sufrido un tipo de delito, personas con discapacidad, etc. Este catálogo serviría para conocer características en particular tales como su número telefónico, su localización, sus características, sus necesidades, etc.

Dentro de estos grupos, hay uno que tiene especial importancia que se relaciona con el tema de los personas con discapacidad, pero no existe un catálogo que permita saber cuántas personas con discapacidad hay en una comunidad, qué necesidades tienen, cuál es su discapacidad, quién los mantiene, a quién mantienen ellos, en qué trabajan, cuánto ganan, etc. Todo esto es muy importante, sin embargo, es muy difícil saberlo en una ciudad.

En el Estado de Colima se tienen 711,235 habitantes [5] y se considera que un 7.4 % de la población padece alguna discapacidad [6]. La forma en que se hacen el conteo tradicional es mediante un censo, lo que significa ir casa por casa verificando si hay personas con alguna discapacidad, a la vez que se realizan entrevistas y se toman datos. Por lo que hacer un catálogo es costoso y lento.

Lo que se propone en esta tesis es aprovechar la estructura de conexión que hay entre las personas con discapacidad, ya que se parte de la suposición de que se conocen entre ellos, debido principalmente a que comparten médicos, asistencia, experiencias y se prestan dispositivos de apoyo, entre otras cosas. Lo que significa que pueden localizarse entre ellos con mayor facilidad. La idea es en lugar de realizar un censo, poder aprovechar la estructura de red, ir con una persona con discapacidad y solicitar que facilite el contacto de donde se encuentren los demás, y así sucesivamente. Pero nuestro problema es el siguiente, ¿cuándo parar? por ejemplo se podría convertir en algo de mayor costo y se tiene el riesgo de no cubrir a todas las personas con discapacidad. Aquí se parte desde la premisa de que la red de personas con discapacidad es del tipo red Watts-Strogatz o red de “Small-Worlds”.[4]

En este trabajo se utilizará el muestreo conocido como “SnowBall” o bola de nieve, el cual emplea un enfoque distinto para estudiar poblaciones difíciles de encontrar o poblaciones escondidas. El muestreo consiste en empezar con un nodo o persona que servirá como semilla, y luego ir a los nodos que están conectados a éste, es decir, a las personas que la persona conoce, después ir con todos los nodos que estén conectados a los nodos anteriores y así sucesivamente. El muestreo se detiene cuando se ha llegado al tamaño de la muestra que se quiere. Es nuestro caso no se detendrá cuando se considera que se ha contabilizado una proporción considerable de individuos de este grupo.

También se intentará aplicar un procedimiento basado en la cobertura de una muestra para decidir si ya se tienen suficientes sujetos muestreados. En otras palabras, se intentará probar una expresión matemática basada en el concepto de cobertura de una muestra, propuesto por Alan Turing y concretado por Good [4], que en teoría permitiría saber o estimar el porcentaje de personas con discapacidad que se ha cubierto con cada muestra, de tal forma que se pueda decidir si se detiene o no.

El índice de cobertura se basa en que en teoría se puede estimar el porcentaje de personas con discapacidad que ya se han detectado, basándose en, el tamaño de muestra y el número de discapacitados que se han nombrado una sola vez mientras se recorre la red, o sea, los nodos a los que se han visitado una sola vez conocidos como singletons.

Capítulo 2

Preliminares

En este capítulo se hará un repaso de definiciones y herramientas que serán de utilidad en este trabajo.

2.1. Redes de Watts-Strogatz

El concepto de la redes de Watt-Strogatz nace del trabajo publicado en el año 1998 por D. Watts y S. Strogatz, del Departamento de Mecánica Teórica y Aplicada de la Universidad de Cornell. Su trabajo “Collective dynamics of ‘small- world’ networks” [1] en la revista *Nature*, en donde exploraron modelos simples de redes regulares “reconectadas” para introducir desorden.

Esas redes las llamaron “Small-Worlds Network” o redes del mundo pequeño, por analogía con el fenómeno del mundo pequeño propuesto por S. Milgram en 1967 [8]. También, los modelos de sistemas dinámicos con acoplamiento de mundo pequeño muestran una velocidad de propagación de la señal mejorada, potencia computacional y capacidad de sincronización.

Las redes de Watts-Strogatz se construyen de la siguiente manera: se comienza a partir

de una red regular o un anillo de N vértices y k aristas por vértices vecinos para cada lado como se puede ver en la Figure 2.1. Por cada arista, con una probabilidad p se conecta esa arista a otro nodo al azar. Entonces tiene los siguientes 3 parámetros:

N : Número de nodos

k : Número de aristas que se conecta a sus nodos vecinos

p : Probabilidad de cambiar la conexión de 2 nodos.

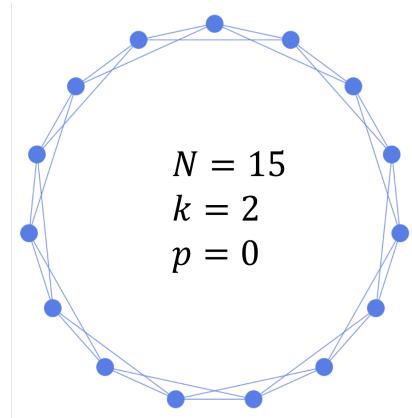
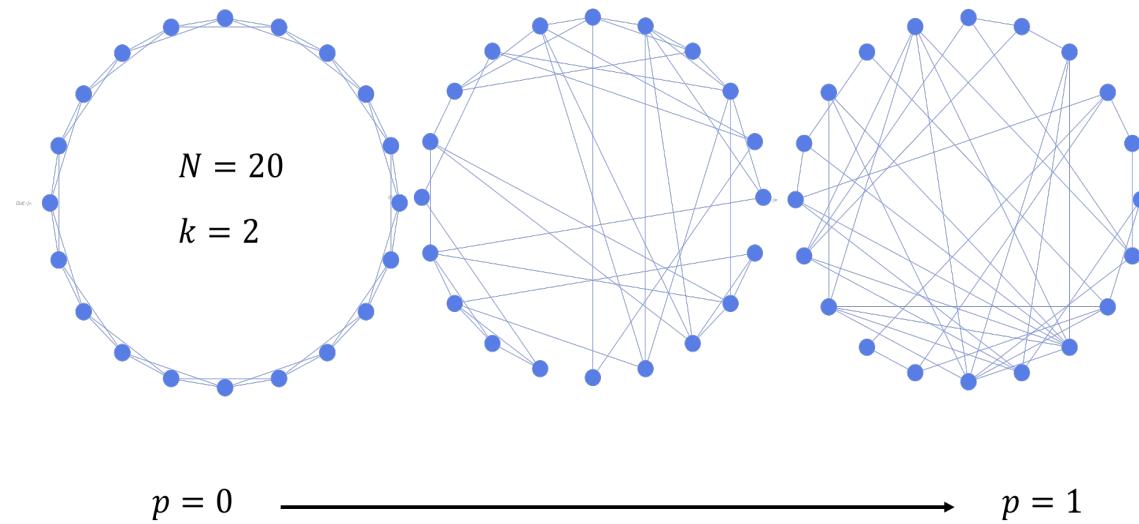


Figura 2.1: Ejemplo de red de Watts-Strogatz.

Cuando p es pequeño se conoce como red de “Small-World” y tiene las propiedades de que la conectividad es alta aunque las personas de la red expresan lo contrario. En la figura 2.2, se puede ver el efecto de la topología de la red mientras $p \rightarrow 1$.

Figura 2.2: Efecto de p en la topología de la red

2.2. Singletons

Los singletons son individuos que solo salen de una vez en la muestra, cuando el número de singletons es alto en una muestra; quiere decir que la variabilidad de las categorías es alta y la muestra es muy pequeña. De manera más sencilla lo podemos observar en la figura 2.5, donde en nuestra muestra solo sale una canica amarilla y una roja, por lo que la amarilla y la roja serían nuestros singletons.

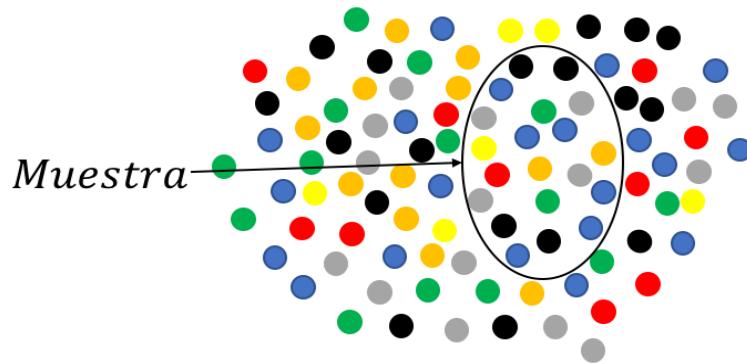


Figura 2.3: Ejemplo de singletons

2.3. Cobertura

La cobertura de una muestra está definida como la proporción de la población de individuos que son representados en una muestra. Si p_i es la proporción de individuos de la clase i en la población, y x_i es el número observado de individuos de clase i en la muestra, entonces C , la cobertura de la muestra es

$$C = \sum_{i=1}^M p_i I(x_i > 0),$$

donde M es el número de clases en la población.

I. J. Good [4] fue el primero en proporcionar una estimación para la cobertura de una muestra en el año 1953, aunque atribuye el resultado principal de su trabajo a la comunicación personal con A. M. Turing.

Varios autores han tratado temas relacionados con la cobertura de una muestra: Good & Toulmin [3] extendieron el trabajo de Good [4] y proporcionaron una estimación al número de especies en una población y analizaron el efecto de una muestra adicional sobre un aumento en la cobertura.

Un ejemplo sencillo se puede ver en la Figura 2.3 y Figura 2.4, donde la muestra se compone de dos individuos de color negro, uno verde y uno azul. Por lo tanto

$$C = \frac{9}{38} + \frac{6}{38} + \frac{6}{38} = \frac{21}{38} \approx 0.55$$

en otras palabras, si se toma otro individuo al azar se tiene un 55 % de probabilidad de que ya se haya muestreado. Esta es la cobertura real.

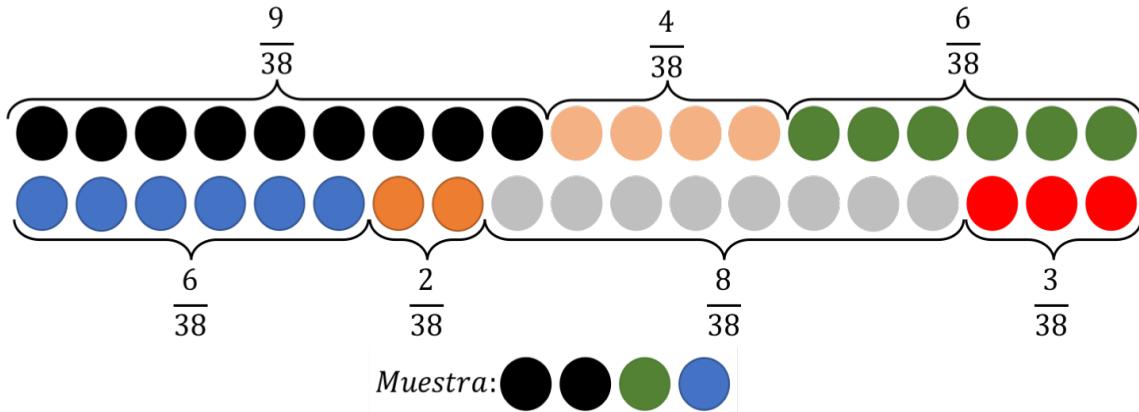


Figura 2.4: Ejemplo de cobertura y muestra del ejemplo de la cobertura

Pero hay un problema, con las poblaciones escondidas o difíciles de alcanzar, para las cuales no se conoce el número total de personas de la población. Para eso se utilizará la siguiente ecuación

$$C \approx 1 - \frac{s}{n} \quad (2.1)$$

donde s son los singletons y n el tamaño de nuestra muestra y se explicará de donde proviene la expresión (2.1).

Hernández [2] menciona que la ecuación (2.1) se puede obtener utilizando modelos de urnas, el cual consiste en una o mas urnas que son llenadas con canicas de diferentes colores que representan a los elementos de interés. Entonces, si se tiene una muestra de n elementos, donde hay x_i individuos de tipo $i = 1, 2, 3, \dots, k$, la pregunta sería, ¿cuál es la estimación por máxima verosimilitud (EMV) de las frecuencias relativas de todas las clases? La función de verosimilitud de una muestra está dada por:

$$P(X = x|p) = \binom{n}{x_1 \ x_2 \ \dots \ x_k} p_1^{x_1} p_2^{x_2} p_3^{x_3} \dots p_k^{x_k} \quad (2.2)$$

que es maximizada cuando $p_i = x_i/n$.

Pero hay un error inherente que resulta ser relevante cuando el número y las frecuencias de las clases en la urna es desconocida. Un ejemplo sencillo para ver esto es lo siguiente: suponiendo que se sacan 3 canicas de diferente color, la urna que maximiza la verosimilitud utilizando (2.2) es que haya $1/3$ de cada color. Sin embargo, esto es verdad si solo hay 3 clases de colores en la urna. Pero si no se sabe los colores que hay adentro de la urna, y se obtiene la misma muestra de 3 colores diferentes, en este caso la composición de la urna que maximizaría la probabilidad de la muestra es que todas las canicas fueran de diferentes colores.

Otro ejemplo es lo siguiente: si se toma una muestra de tamaño 10 y se obtiene 4 canicas negras, 3 blancas, 1 roja, 1 azul y 1 amarilla. El estimador clásico para esos respectivos colores es:

$$p = \{4/10, 3/10, 1/10, 1/10, 1/10\}$$

La verosimilitud de la muestra es 1.7×10^{-2} , pero el problema aquí es que no se sabe nada sobre la composición, o el número de colores en la urna. La pregunta sería ¿qué composición de urna maximiza la probabilidad de sacar 4 negras, 3 blancas, 1 roja, 1 azul y 1 amarilla?. Aun mejor, ¿cuál composición de urna maximiza la probabilidad de conseguir canicas en las frecuencias observadas?. Para responder esto se necesita tener cuidado, porque si hay singletons en la muestra, existe otra composición de la urna que tiene más alta verosimilitud que usar la estimación tradicional $p_i = x_i/n$.

Si se regresa al ejemplo anterior, se mostrará que existe una composición de una urna tal que su EMV es 27 veces más grande que usando la composición tradicional. Para esto se asume que la muestra viene de una urna que contiene N canicas que se dividen en M clases y los dos son desconocidos. Asumiendo que hay K clases que contiene más de un individuo cada uno , entonces $S = M - K$ son el número de clases que solo tienen un único individuo cada una. Sea θ la fracción de la población que es ocupada por las S clases.

Suponemos que la muestra es de tamaño n y contiene m diferentes clases, de las cuales s son singletons. Posteriormente, nombramos las clases arbitrariamente como $1, 2, 3, \dots, m$, y sea x_i el número de individuos en la muestra perteneciente a i . La verosimilitud de la muestra es:

$$P(X = x|p) = \binom{n}{x_1 \ x_2 \ \dots \ x_{m-s}} p_1^{x_1} p_2^{x_2} p_3^{x_3} \dots p_{m-s}^{x_{m-s}} p_{m-s+1}^1 p_{m-s+2}^1 \dots p_m^1 \quad (2.3)$$

donde p_i es la proporción en la población representado por los x_i individuos en la muestra. Se observa que en el tradicional EVM para distribuciones multinomiales, la dimensión de los parámetros es restringida al número observado de clases. Sin embargo, si la dimensión del espacio de parámetros no es restringida al número conocido de clases, el producto de los últimos s de p_i en (2.3), es decir

$$p_{m-s+1}^1 p_{m-s+2}^1 \dots p_m^1$$

es de hecho la probabilidad de conseguir s diferentes individuos con la muestra de tamaño s . Esta probabilidad es maximizada cuando la muestra de tamaño es tomado fuera de la fracción θ de la población, cuyos individuos pertenecen a diferentes clases. Entonces, la verosimilitud se puede reescribir como:

$$P(X = x|p) = \binom{n}{x_1 \ x_2 \ \dots \ x_{m-s}} p_1^{x_1} p_2^{x_2} p_3^{x_3} \dots p_{m-s}^{x_{m-s}} \theta^s$$

Así, el EMV es

$$\hat{p}_i = x_i/n, \quad \hat{\theta} = s/n$$

lo cual da una verosimilitud de 4.7×10^{-1} . Por lo tanto, la oportunidad de obtener una muestra que coincide con la observada es mayor si la urna contiene 4/10 negras, 3/10 blancas y que el resto sean canicas totalmente de diferente color. Bajo estos argumentos dados, en (2.3) los colores no detectados en la muestra deben pertenecer a la fracción θ

para los cuales cada individuo pertenece a diferente clase. Entonces, un estimado de la fracción de la población que no ha sido representada en la población es precisamente $\hat{\theta}$ menos la fracción de s/N correspondiente a las frecuencias poblacionales de los singltons detectados en la muestra, es decir, $\hat{\theta} = s/N$. Por lo tanto, el EMV de la cota inferior para la cobertura de la muestra es

$$\hat{C} = 1 - s/n$$

el cual es precisamente la estimación de Good.[3]

Para el ejemplo de la Figura 2.3 y Figura 2.4, obtenemos que $C \approx 1 - s/n = 1 - 2/4 = 0.5$, y se puede ver que se acerca a la cobertura real, que es 0.55. Para la Figura 2.5, $C \approx 1 - 2/20 = 9/10 = 0.9$, lo cual significa que tenemos un 90 % de que el próximo individuo que se muestrea ya lo hayamos observado antes. En este ejemplo la cobertura exacta es 1.

2.4. Diseños de Link-Tracing y muestreo de bola de nieve (SnowBall Sampling)

D. D. Heckarthon y C. J. Cameron [7] mencionan que los diseños de Link-Tracing tienen origen en Bureau of Applied Social Research en Columbia en el año de 1940, debido a que los métodos tradicionales de encuesta no eran adecuados para estudiar la relación entre las opiniones de los líderes y los seguidores, se les preguntó el nombre de las personas que los influenciaron. De ahí, la segunda ola de personas fueron entrevistadas. Coleman [9] en 1958 y Goodman [10] en 1961 estaban estudiando la estructura de redes sociales. Iniciaron con una muestra aleatoria que las llamaron semillas, la muestra se expandió ola por ola para revelar los patrones de influencia entre los médicos, y otros fenómenos.

Muchos años después de el trabajo de Coleman y Goodman, una nueva aplicación de

esos diseños emergió como un enfoque no probabilístico para estudiar poblaciones escondidas o difíciles de encontrar. Muestras de poblaciones escondidas deben iniciar con una muestra fija de personas iniciales, porque si fuera al azar entonces no sería una población escondida en primer lugar. Estos sujetos iniciarián como semillas a través del cual se reclutan una ola de personas, una ola de personas que reclutan otra ola de sujetos, y luego la muestra se expande ola a ola como una bola de nieve crece en tamaño a medida que rueda cuesta abajo. A este muestreo se le llama “muestro de bola de nieve”.

Capítulo 3

Técnica de muestreo de redes utilizando generaciones

3.1. Hipótesis de la red de la población y de la subred del grupo difícil de encontrar.

¿Cómo se puede conocer el porcentaje de la población de personas que se lleva muestreado?. Para tomar en cuenta este puntos se harán varias suposiciones sobre la población que mas adelante se utilizarán. Primero, la red de la población tiene que satisfacer que las conexiones sean bidireccionales, es decir, las personas se conocen mutuamente. Dentro de esta red de la población debe existir una subred conectada, en este caso, la red de personas; esto tiene sentido porque de otra manera no se podría encontrar a todas las personas en la subred.

El problema consiste en encontrar un método eficiente para encontrar todos los elementos de la subred de personas sin recorrer toda la red de la población aleatoriamente. Por lo que se asumirán dos cosas, la primera es que al menos se conoce una persona o un nodo de la subred, porque de otra manera no se puede iniciar; la segunda es que la subred

de personas discapacitadas es una red de Watts-Strogatz.

Notación: La red de la población y la subred de personas con discapacidad se denotarán por H y G , respectivamente.

3.2. Descripción de la técnica

En esta técnica se utilizarán generaciones de personas, y cada generación será la que se tomará como muestra. En cada muestra o generación se calculará los singletons y en base a eso se usará la expresión matemática (2.1) de la cobertura.

La primera generación será el nodo o persona que se conozca por la hipótesis, la segunda generación serán todos los nodos o personas a los que nos lleva el nodo que se conoce, la tercera generación serán todos los nodos o personas a los que nos llevan cada una de las personas que estuvieron en la segunda generación (generación pasada) y así sucesivamente. Debido a que en cada generación se están calculando singletons y la cobertura, esta técnica se deja de hacer cuando se terminan los singletons.

Esta técnica se puede visualizar en la Figura 3.1, donde se muestra la red H y la subred G . En este caso, se empieza con el nodo 4, que será la primera generación, por lo que el número de singletons y el tamaño de muestra son iguales a uno. En la figura 3(e) se muestra la segunda generación, señaladas con flechas de color negro. Después se repite el procedimiento anterior con cada uno de los nodos de la segunda generación, y se calcula los singletons y el tamaño de muestra para obtener la cobertura estimada. En la Figura 3.2 se puede ver la continuación de la técnica que se repite hasta que ya no hay singletons.

La ventaja de esta técnica es que se abarca mucho en pocas generaciones, pero en cuestiones prácticas es tardado ir a todos los nodos a los que manda la generación actual.

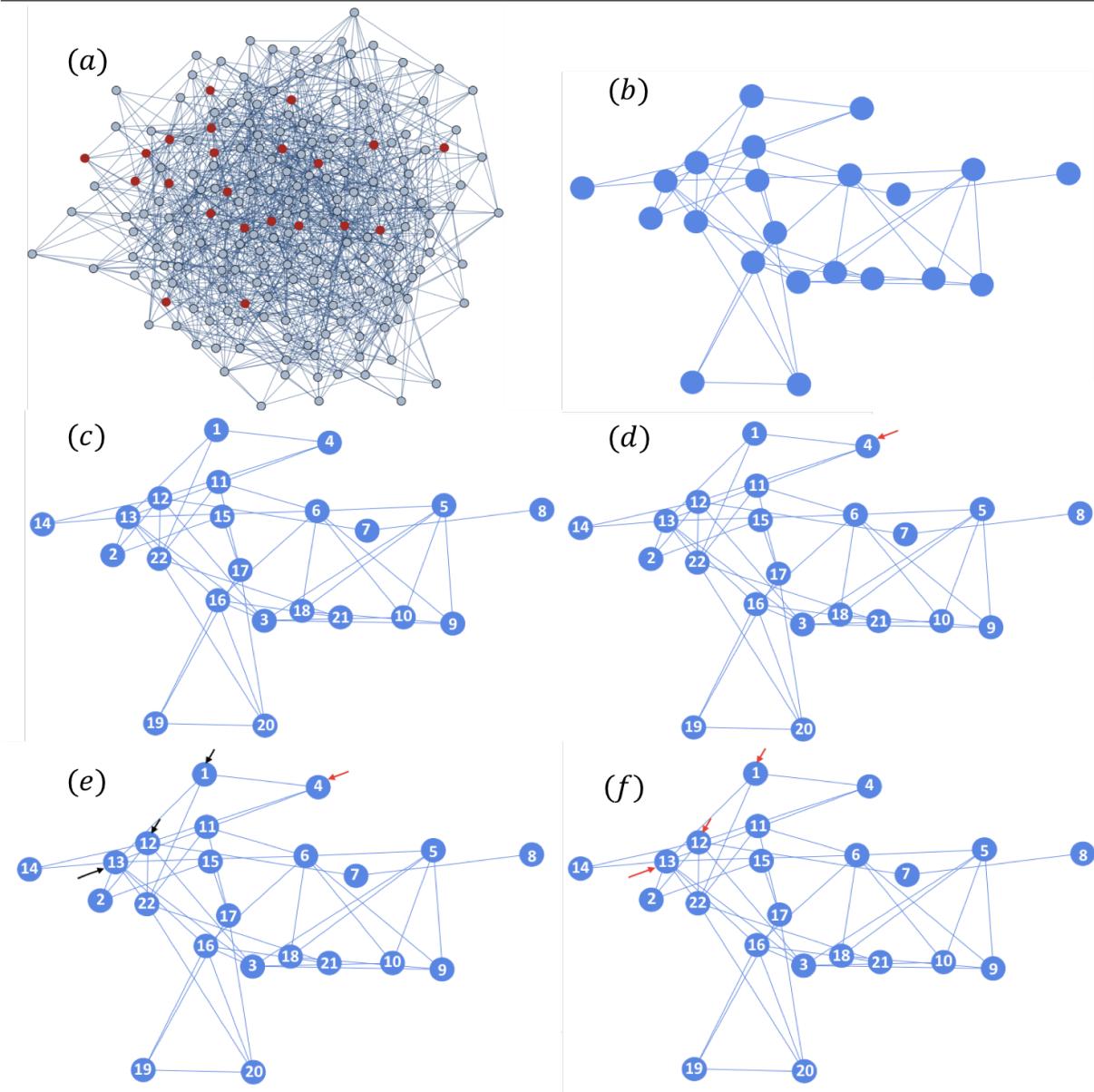


Figura 3.1: (a) Red H , (b) Subred G , (c) G enumerada, (d) Nodo que conocemos, (e) Nodos que conocen el primero, (f) Nodos nuevos

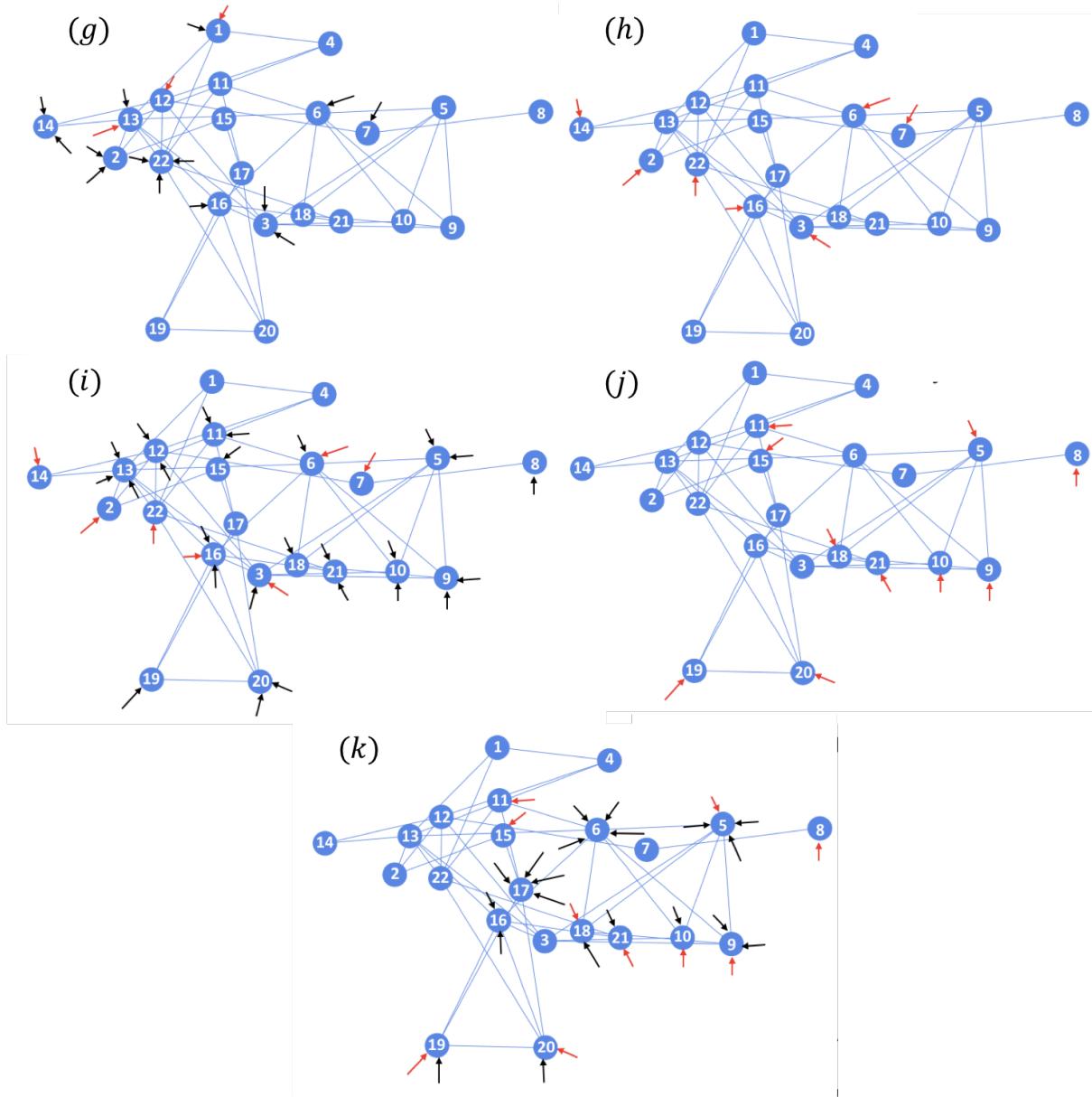


Figura 3.2: (g) Nodos que conocen los nodos nuevos, (h) Nodos nuevos, (i) Nodos que conocen los nodos nuevos (j) Nodos nuevos, (k) Nodos que conocen los nodos nuevos

Obteniendo el número de singltons en cada muestra y el tamaño de muestra; se acumulan con cada generación, se puede obtener la cobertura. Graficando los datos del ejemplo de la cobertura real y la estimada se obtiene la Figura 3.3.

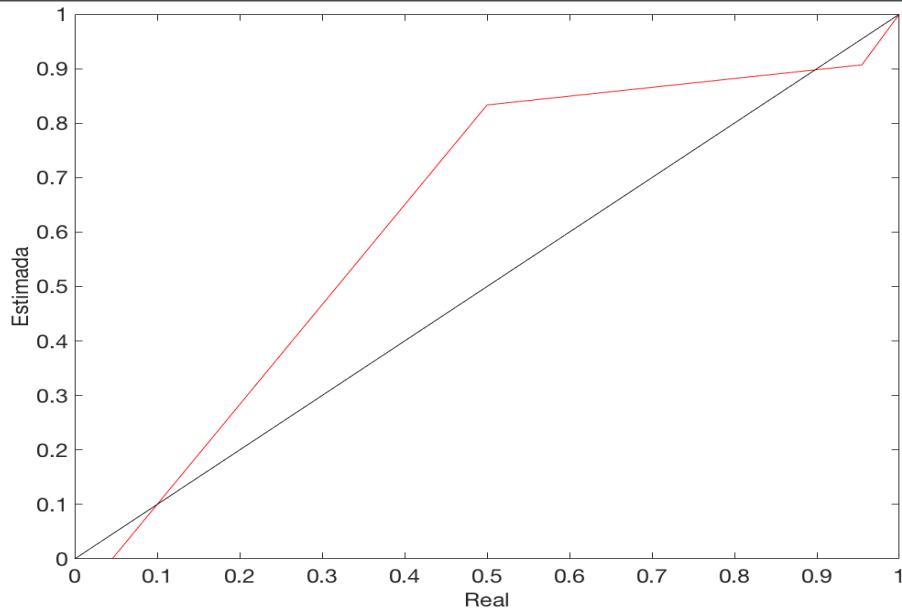


Figura 3.3: Grafica de real contra estimada

En la gráfica, el eje x es la cobertura real, el eje y es la cobertura estimada, y la línea negra es la función $y = x$, para que este método sea eficiente, la curva de color rojo tiene que acercarse a la recta negra, en otras palabras, la cobertura estimada se aproxima a la real. Si la curva de color roja está por arriba de la línea negra significa que la cobertura estimada está sobreestimada y si pasa por abajo significa que es subestimada.

3.2.1. Simulaciones

Para ver el comportamiento de la cobertura estimada utilizando la técnica anteriormente descrita, se hicieron simulaciones en Matlab. Se realizaron 10 simulaciones utilizando una población de 1000 personas, el coeficiente definido en preliminares, $k = 3$ y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$, los resultado se muestra en la Figura 3.4.

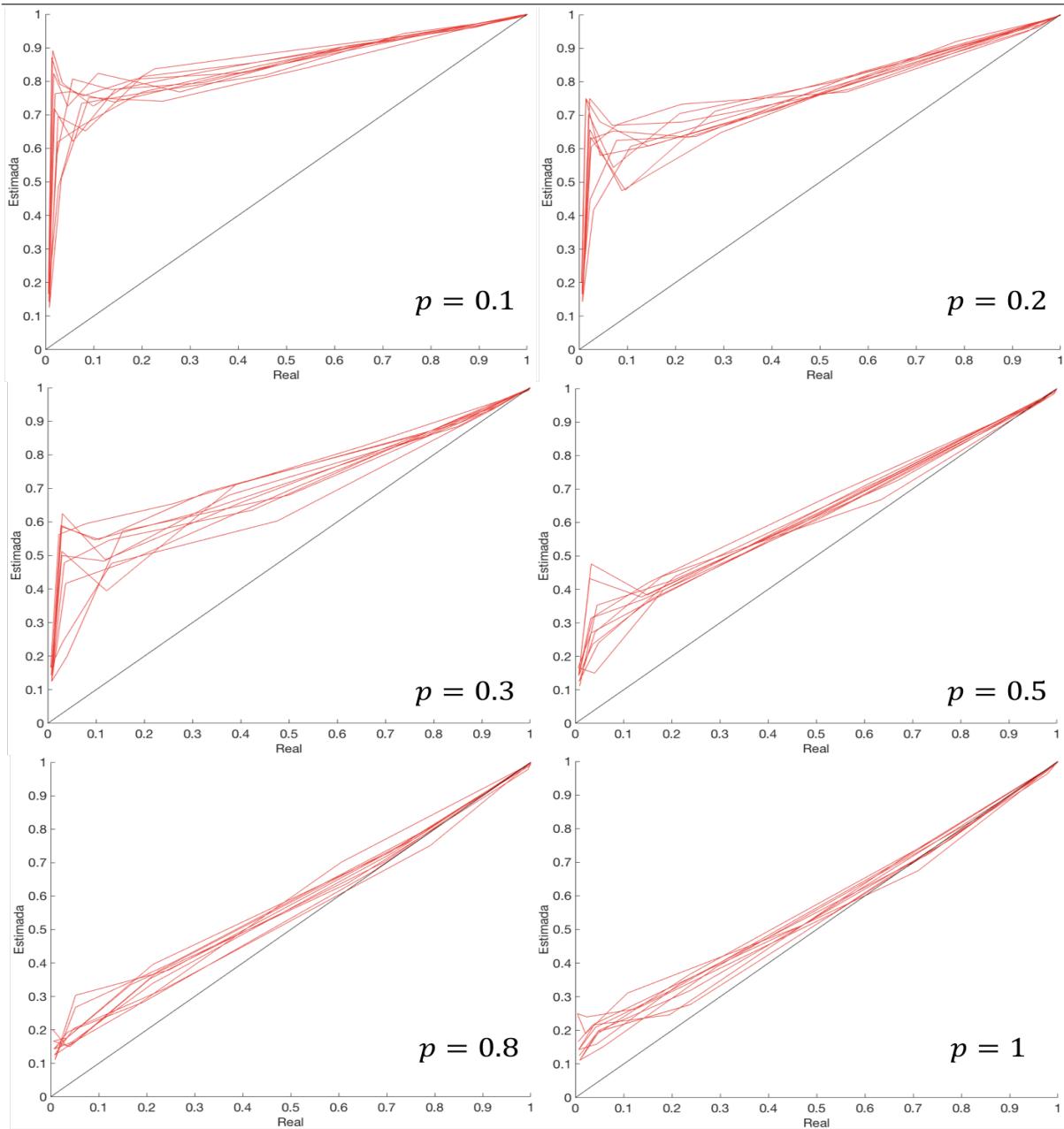


Figura 3.4: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$.

3.2.2. Observaciones

A partir de las simulaciones se puede notar que cuando se llegue a cobertura muy alta, la aproximación será similar a la real, pero no es problema, porque es justamente lo que se quiere y mientras $p \rightarrow 1$ el error tiende a 0.

Se puede notar que la línea estimada pasa muy arriba de la real cuando p es pequeño, lo que significa que el tamaño de la muestra sigue siendo muy grande en cada generación, siendo una aproximación no tan precisa. Para mejorar el esquema de muestreo se propone un término lo que se llama *Lag*, que ayudará a quitar elementos de la muestra.

3.3. Mejora de la técnica agregando el término Lag

El *Lag* se utilizará para eliminar elementos y reducir el tamaño de la muestra. Se retiran los elementos que hayan salido generaciones anteriores a la actual, si se quita los elementos que estén en una generación anterior de la generación actual se denominará como Lag_1 . De hecho, la técnica anterior es un método que se usa una especie de Lag_1 debido a que solo quita el nodo de quien mandado. En general, si se quitan los elementos de k generaciones anteriores sería Lag_k .

Si se denota al conjunto S_n como los elementos que están en la generación n , y LS_n^k como los elementos que están en la generación n con Lag_k

$$x \in LS_{n+1}^k \quad \text{si} \quad x \in S_{n+1} \quad \text{y} \quad x \notin \bigcup_{i=n-k}^{n-1} S_i$$

Si $n - k < 0$, entonces i inicia en 0.

3.3.1. Simulaciones

Para ver el comportamiento de la cobertura, se hicieron 10 simulaciones con $N = 1000$, $k = 3$, $Lag = 1$ y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$, los resultados se muestran en la

la Figura 3.5. Además 10 simulaciones con $N = 1000$, $k = 3$, $Lag = 10$ y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$, los resultados se muestran en la Figura 3.6

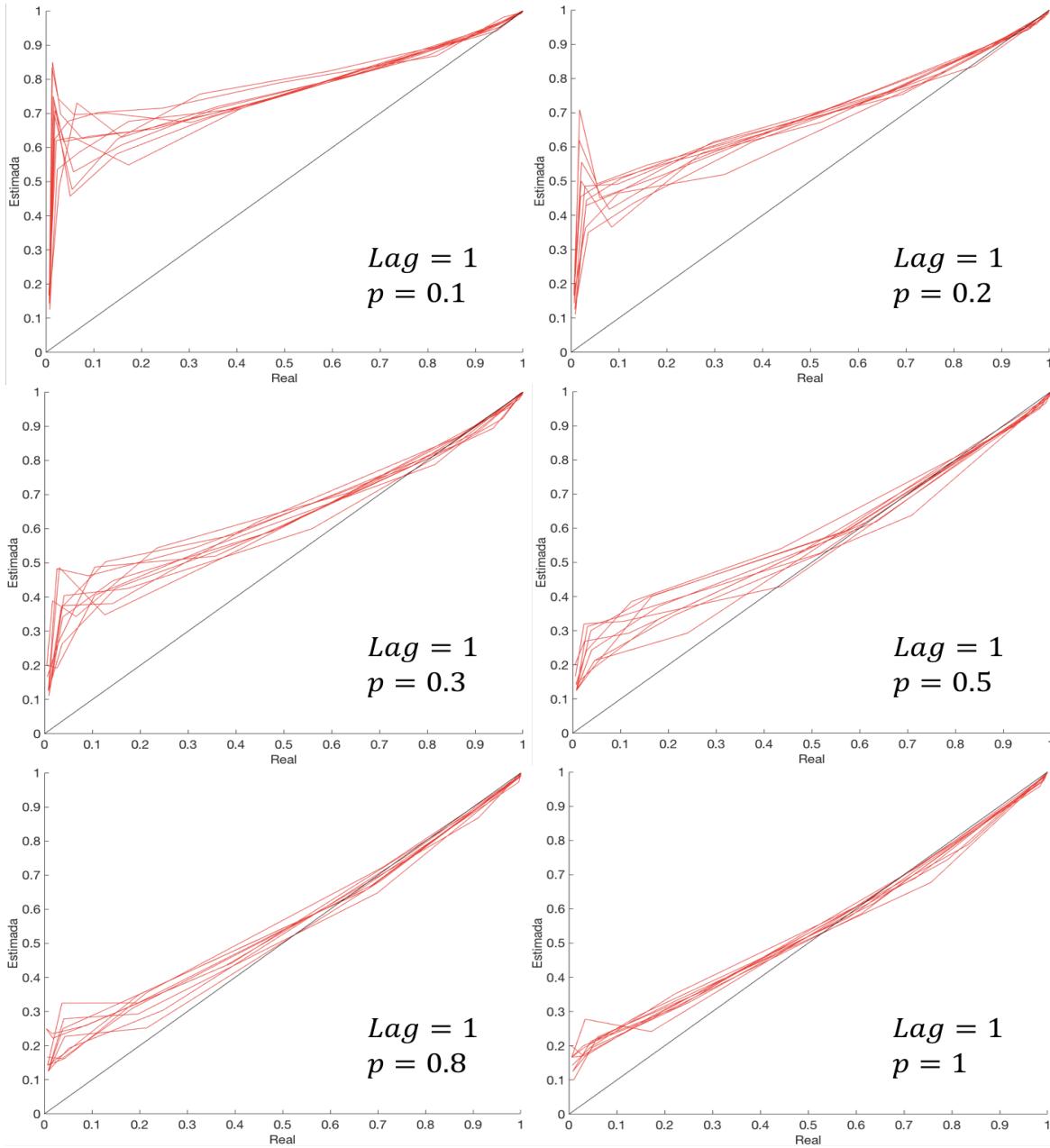


Figura 3.5: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$ utilizando $Lag = 1$

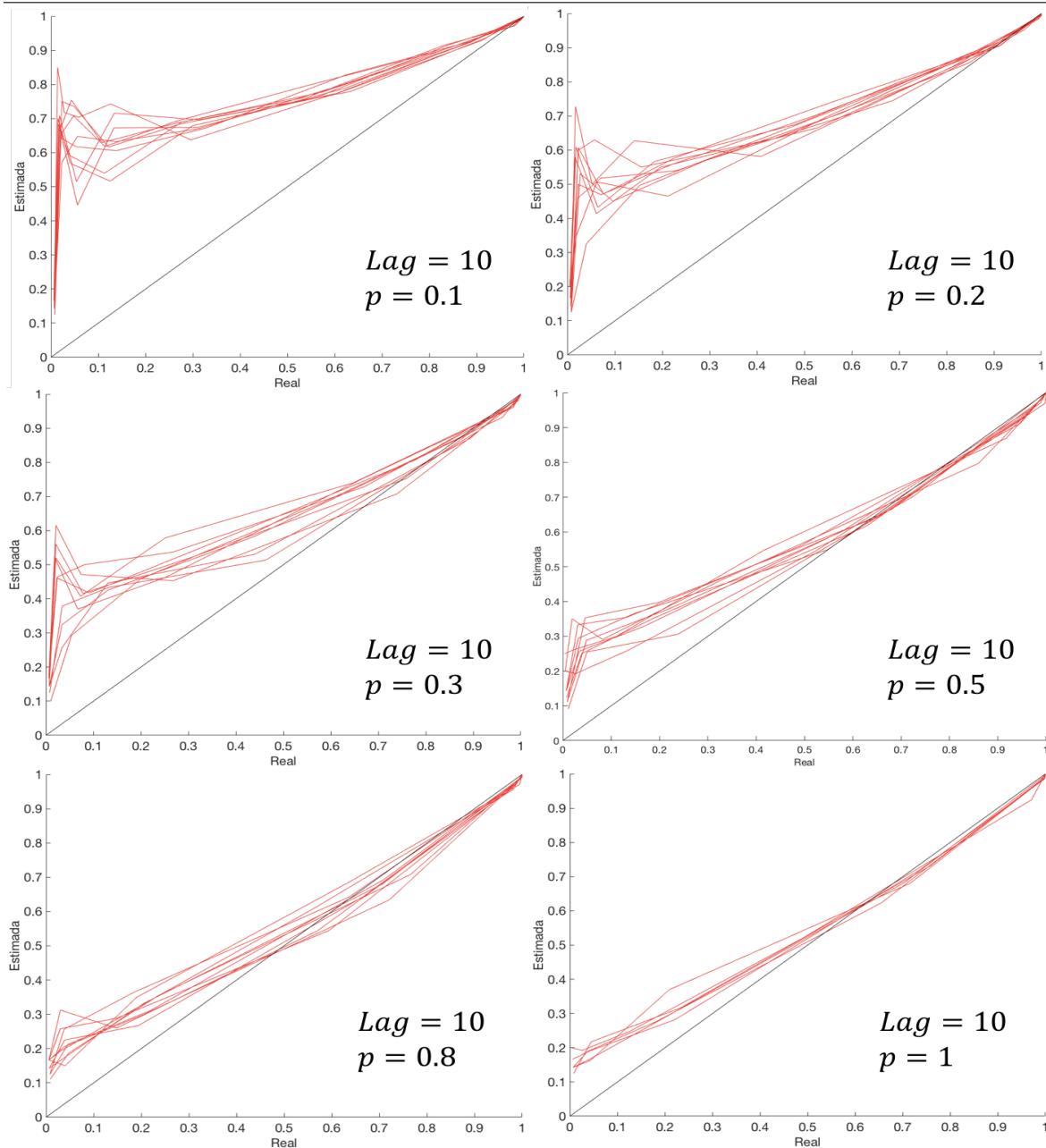


Figura 3.6: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$ utilizando $Lag = 10$

3.3.2. Observaciones

Claramente hay una mejora con respecto al método sin *Lag*, las líneas rojas (estimada) ya se acercan más a la línea negra (real), pero el caso que queremos atacar es cuando $p = 0.1$, que aún sigue siendo una mala aproximación, debido la cobertura real y estimada se parecen hasta tener un cobertura estimada de más de 94 %, por lo que el tamaño de la muestra sigue siendo muy grande.

¿Se puede mejorar esta aproximación?, ¿qué pasará si en la generación actual se quitan todos los elementos de todas las generaciones anteriores a la actual?. La respuesta es sí, sí se puede mejorar, para esto es necesario quitar todos elementos de todas generaciones para que la muestra no sea demasiado grande.

3.4. ¿Qué pasa con Lag_∞ ?

El Lag_∞ es quitar todos los elementos de todas las generaciones anteriores, es decir, para calcular los singletons y la cobertura en la generación actual, solo se tomarán en cuenta los nuevos nodos que no hayan salido en todas las generaciones anteriores a la generación actual. Este *Lag* se puede ver de la siguiente forma,

$$x \in LS_{n+1}^\infty \quad \text{si} \quad x \in S_{n+1} \quad \text{y} \quad x \notin \bigcup_{i=0}^{n-1} S_i$$

en palabras simples, solo importan los elementos de las nuevas generaciones. Además, la muestra se tomará como los nodos diferentes que se tengan muestreados.

3.4.1. Simulaciones

Para ver el comportamiento de la cobertura estimada, se hicieron 10 simulaciones con $N = 1000$, $k = 3$ y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$, y los resultados se pueden ver en la Figura 3.7

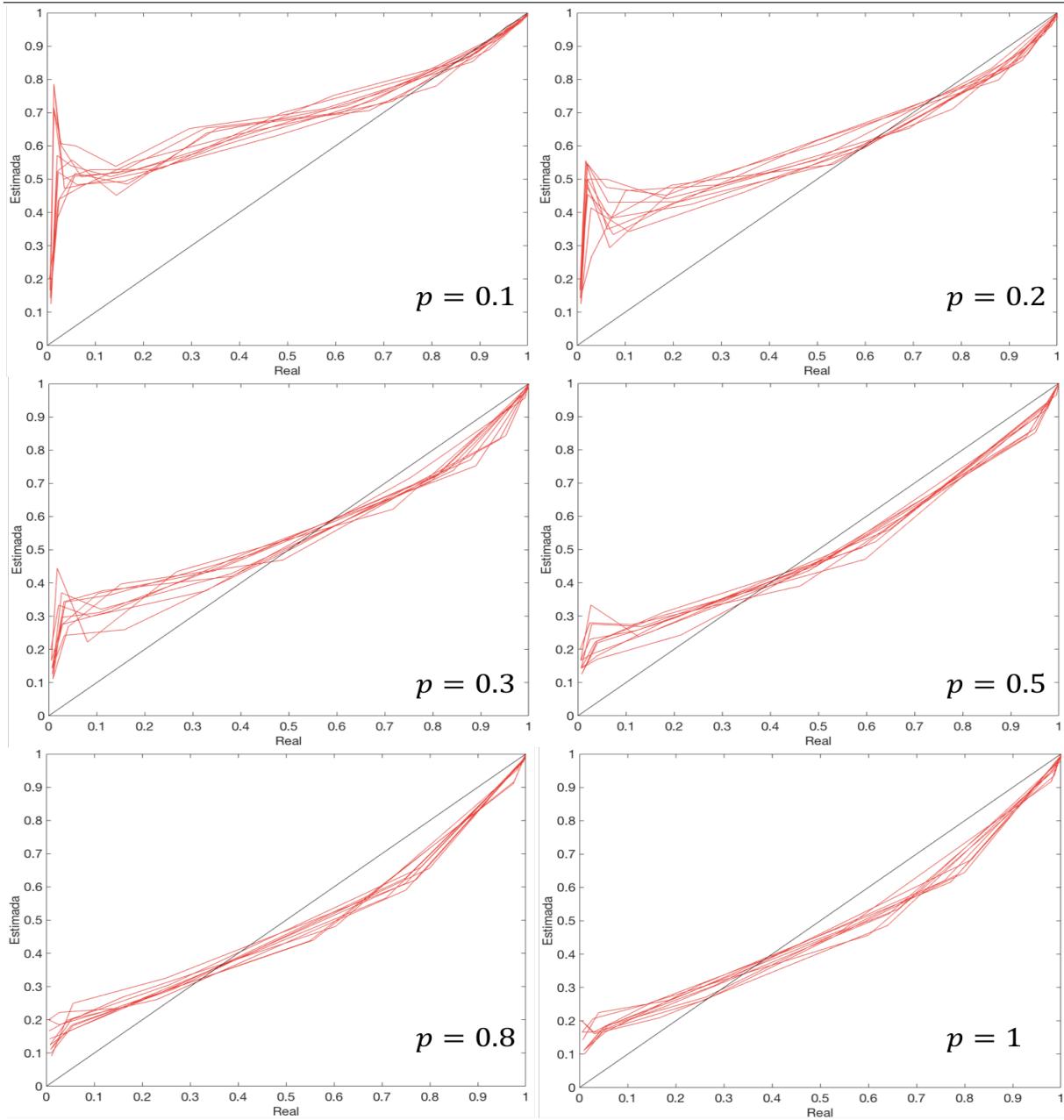


Figura 3.7: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 1000 nodos con coeficiente $k = 3$

3.4.2. Observaciones

Se puede notar que cuando $p = 0.1$, la cobertura estimada se acerca a la cobertura real cuando se lleva al menos 80 % de cobertura, lo cual es una mejora grande respecto al anterior, y además aquí se tiene un resultado conservador, debido que ahora se encuentra por debajo de la cobertura real. Esto en la práctica es bueno, porque se sabe que la cobertura real a lo menos será la cobertura estimada, entonces cuando se tenga un 80 % ya se puede decir con certeza que falta poco para terminar sin importar el p que se tenga.

Por lo que se queda con la técnica de Lag_{∞} , que es la que mejor aproxima la cobertura real desde el 80 % aun estando en el peor de los casos.

Capítulo 4

Técnica de muestreo utilizando caminata aleatoria

Para esta técnica se asumirán las mismas hipótesis del capítulo 3.1. En esta técnica difiere a las anteriores porque no se usará generaciones, a cambio de eso se calcularán los singletons de toda la muestra que se vayan obteniendo y no solo de la nueva muestra, es decir, que ahora se quiere que la muestra que se vaya acumulando no haya singletons y en base a eso se usará la expresión matemática de la cobertura.

4.1. Descripción de la técnica

Se empieza con el nodo o persona que se conoce por hipótesis, esta será la muestra, después se irá con los nodos que están conectados al nodo inicial y la muestra serán los nodos nuevos más el nodo inicial. De la muestra se calcularán los singletons y la cobertura. El siguiente paso será elegir un nodo al azar de los nodos nuevos, y repetir lo mismo, por lo que ésta se deja de hacer cuando se terminan los singletons.

De aquí pueden surgir algunos detalles. ¿Qué pasa cuando en los nuevos nodos ya hayan salido antes? ¿Cómo se elige el siguiente nodo al azar? ¿Se detiene el proceso? Para esto se tiene que tomar en cuenta lo siguiente, sea NA el conjunto de nodos que son elegidos al azar para continuar el muestreo, cuando los nuevos nodos ya estén en la muestra se tiene que verificar si los nuevos nodos también están en el conjunto NA ; si hay alguno que no esté ahí, entonces se elige ese nodo para seguir continuando el muestreo. Pero si los nuevos nodos ya están en la muestra y también en el conjunto NA , además si el número de singleton no es cero, entonces se elige un nodo al azar de los nuevos y se continua con la técnica.

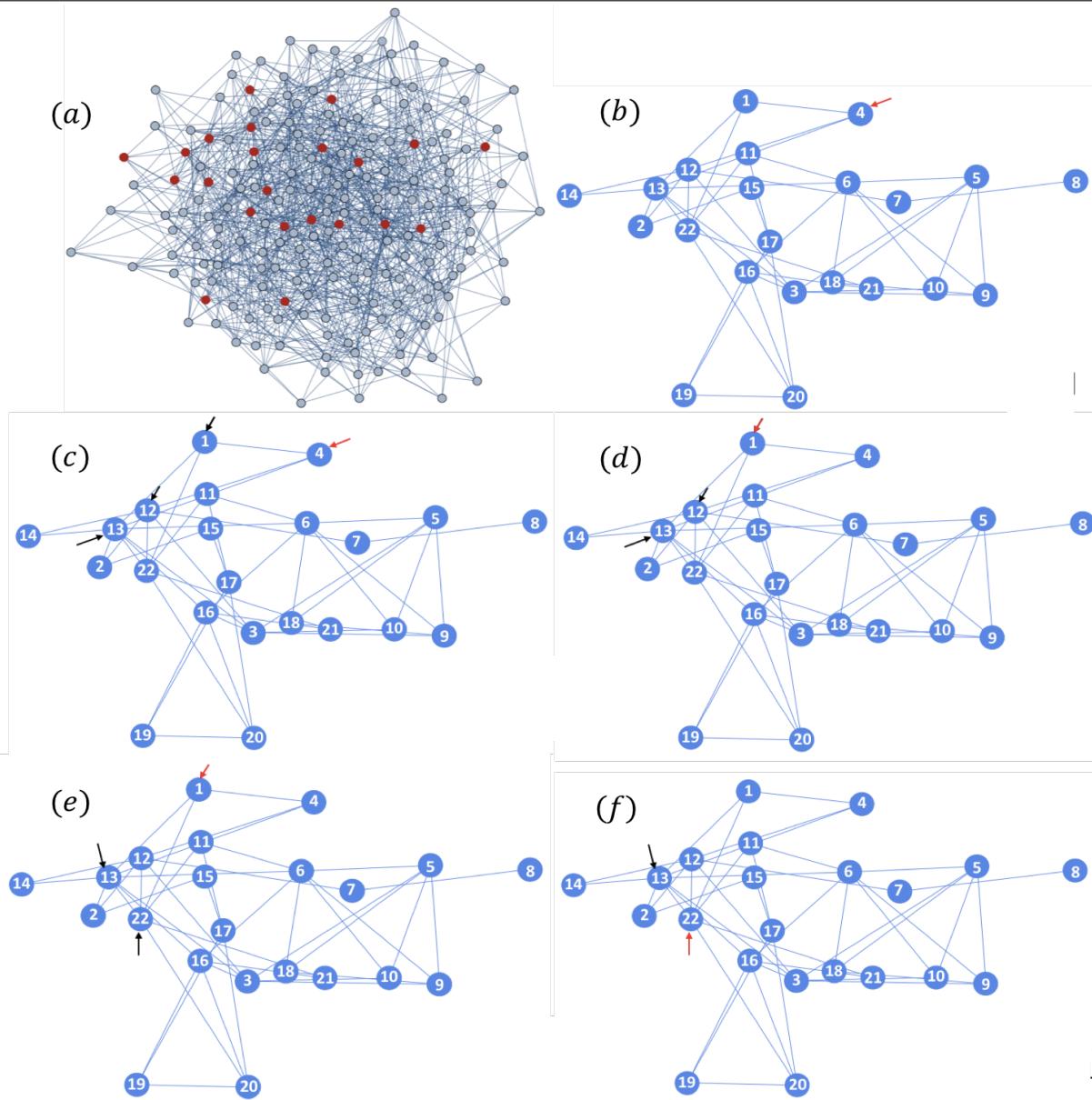


Figura 4.1: (a) Red H, (b) Subred G con el nodo que conocemos, (c) Nodos que se conectan con el primero, (d) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (e) Nodos que se conectan al nodo elegido al azar, (f) Nodos nuevos y nodo con flecha roja es el que se elige al azar

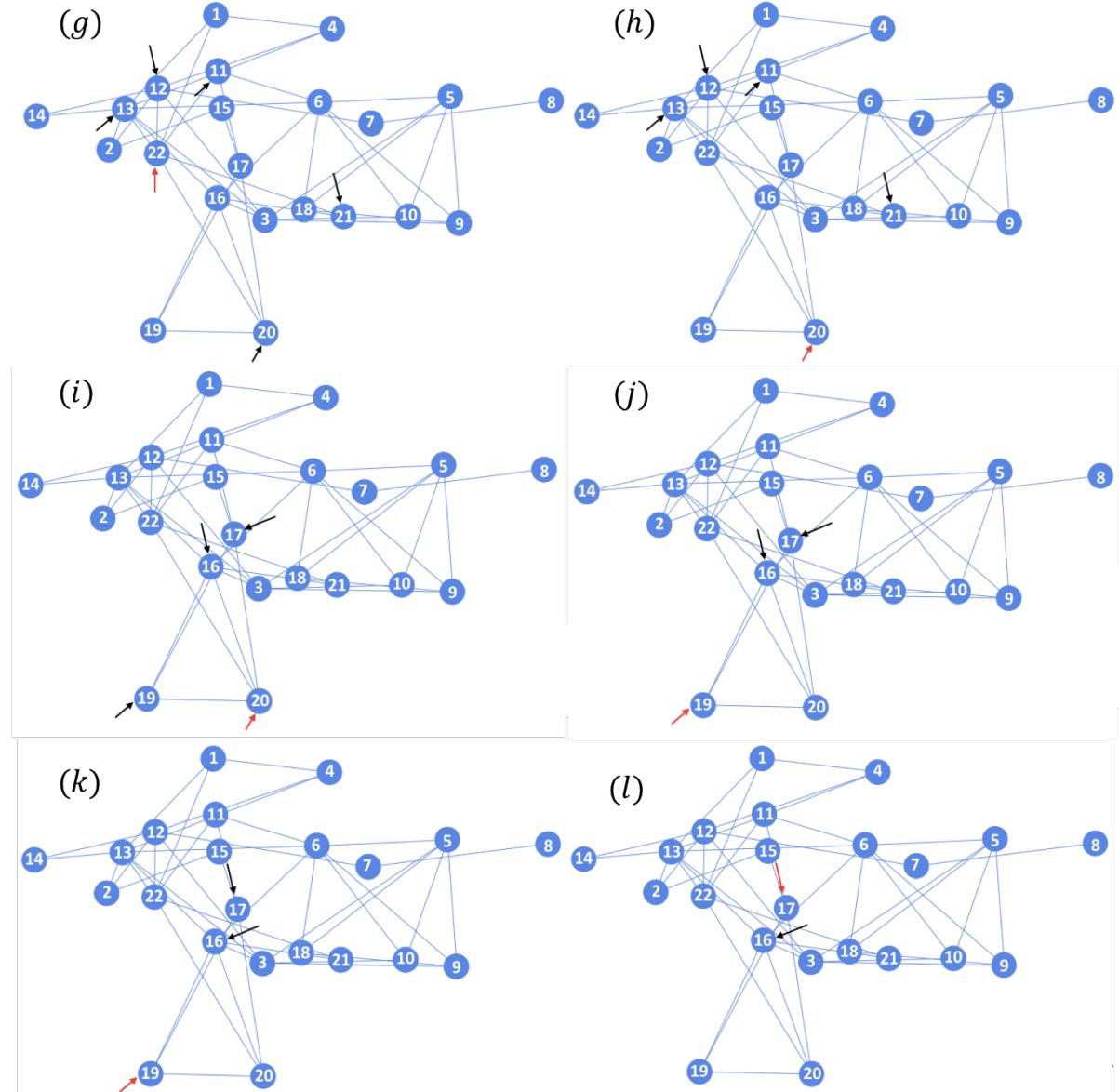


Figura 4.2: (g) Nodos que se conectan al nodo elegido al azar, (h) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (i) Nodos que conocen el nodo elegido al azar, (j) Nodos nuevos y nodo con flecha roja es el que se elige al azar, (k) Nodos que conocen el nodo elegido al azar, (l) Nodos nuevos y nodo con flecha roja es el que se elige al azar

4.1.1. Simulaciones

Para ver el comportamiento se hicieron 10 simulaciones en Matlab con $N = 100$, $k = 3$, y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$. Aquí se bajó el tamaño de la población debido a que si se utilizaba la misma cantidad de personas que en los anteriores se tardaría mucho por la limitaciones del hardware utilizando, por lo cual se tomó esta decisión.

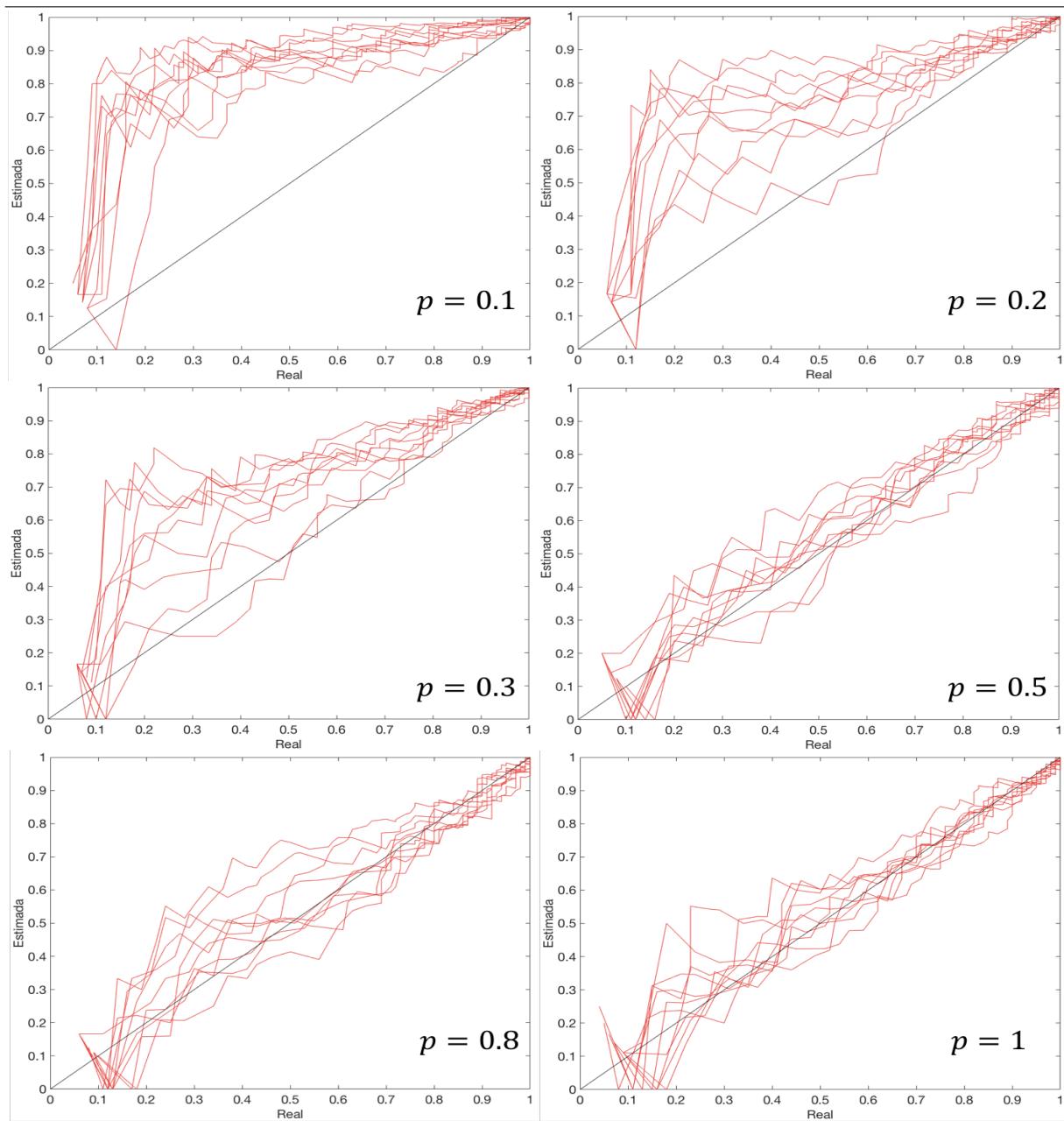


Figura 4.3: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 100 nodos con coeficiente $k = 3$

4.1.2. Observaciones

De la simulaciones se puede ver que la cobertura estimada pasa por encima de la cobertura real para $p = 0.1, 0.2, 0.3$, y hasta $p = 0.5$ ya alcanza la cobertura real. Lo que quiere decir que la muestra es muy grande, esto debido a que cuando los nuevos nodos ya están en la muestra pero no están en NA , y aún quedan singletons; dicha muestra sigue creciente pero los singletons son los mismos, por lo que la cobertura estimada crece. ¿Se podrá mejorar?, ¿qué pasa ahora si no se toma el tamaño completo de la muestra, sino solo el número de nodos que se hayan muestreado?.

4.2. Mejora de la técnica utilizando la muestra como el número de nodos muestreado

Lo que ahora se propone para la cobertura estimada, es que en vez de dividir por el número de la muestra, sea por el número de elementos sin repetición de la muestra, es decir, el número de nodos que se hayan muestreado y así poder reducir el tamaño de la muestra; además de que cambia el tamaño de la misma y no cambia cuando los nuevos nodos ya están en la muestra y también están en NA , y aún quedan singletons

4.2.1. Simulaciones

Para ver el comportamiento se hicieron 10 simulaciones en Matlab con $N = 100$, $k = 3$, y variando $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$

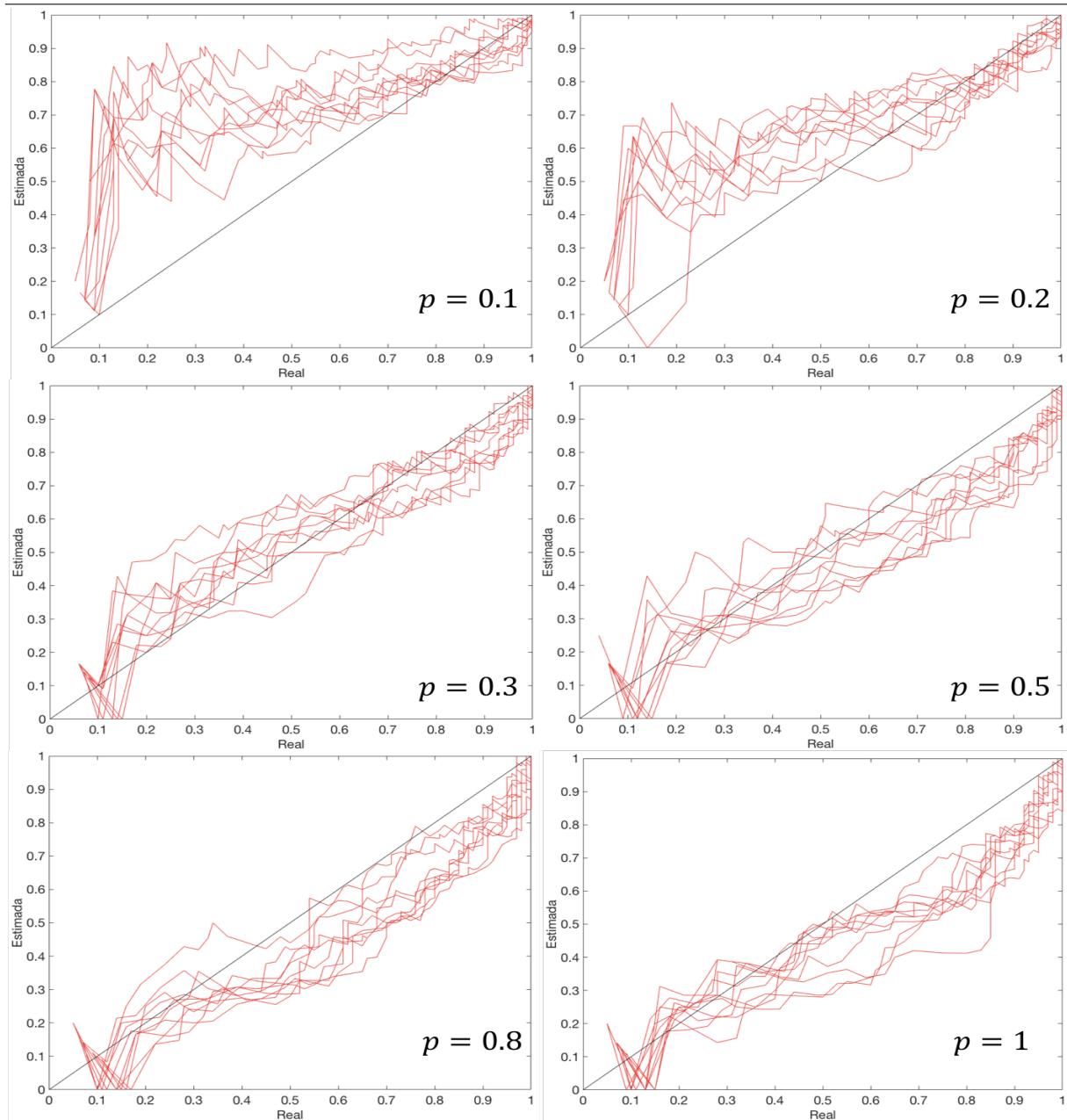


Figura 4.4: Cobertura estimada contra cobertura real para $p = 0.1, 0.2, 0.3, 0.5, 0.8, 1$ correspondiente a una red de 100 nodos con coeficiente $k = 3$

4.2.2. Observaciones

Como se puede observar en la Figura 4.4, esta mejora con respecto a la técnica anterior, pues cuando se está con $p = 0.1$, la cobertura estimada se acerca a la real cuando se alcanza a un 90 % de cobertura, además de tener un resultado conservador cuando $p > 0.4$.

Por lo que se decide quedar con la última técnica, que es la mejor se aproxima a la cobertura real.

Capítulo 5

Conclusiones

En las dos técnicas propuestas y mejoradas en este trabajo de tesis resultaron ser herramientas efectivas para saber el porcentaje de personas o nodos que se llevan en la muestra y además señalar cuánto porcentaje de personas o nodos hacen falta para terminar. Cuando se llega a un 90 %, se puede decir con certeza que es la cobertura real o al menos una cota inferior de la cobertura real, lo cual es un resultado conservador pero eficiente al momento de realizarlo en la búsqueda de personas. Cabe recalcar que este método no solo se limita a personas con discapacidad, sino también a poblaciones escondidas o difíciles de buscar.

Acorde con nuestros resultados, la técnica de muestreo utilizando generaciones es más eficiente que la técnica de muestreo utilizando caminata aleatoria, debido a lo visto en las gráficas; en la primera técnica se sabe que nuestra cobertura estimada es igual o menor a la real cuando se tenga al menos un 80 %, mientras que en la segunda técnica es de un 90 %; esto claramente hablando en el peor de los casos que es cuando $p = 0.1$. Pero la ventaja de la segunda técnica es cuando solamente podemos ir a un nodo a la vez.

Por otro lado, quedan algunas preguntas abiertas, una de ellas es, teniendo ya muestreado la población ¿cómo se puede saber que p le corresponde?. La idea para verificarlo

sería en que momento la línea roja de las gráfica cruza a la recta $y = x$ (línea negra). También falta ver como afecto de N y k , duración en término de generaciones del muestro y utilizar otro tipo de red que no sean Watts-Strogatz.

Bibliografía

- [1] WATTS, D. J. & STROGATZ, S. H., (1998), *Collective dynamic in ‘small-world’ networks*, Nature 393:440-442.
- [2] HERNANDEZ-SUAREZ, C., (2018), *Measuring the representativeness of a germplasm collection*, Biodiversity and Conservation 27:1471-1486.
- [3] GOOD, I. J., (1953), *The population frequencies of species and the estimation of population parameters*, Biometrika 40:237-264.
- [4] GOOD, I. J. & TOULMIN, G., (1956), *The number if new species, and the increase in population coverage, when a sample is increased*, Biometrika 43:45-63.
- [5] INEGI, (2019), *Número de habitantes, Colima* <http://cuentame.inegi.org.mx/monografias/informacion/col/poblacion/>
- [6] INEGI, (2019), *La discapacidad de México, datos 2014* http://internet.contenidos.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825090203.pdf
- [7] HECKARTHON, D. D. & CAMERON, C. J., (2017), *Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling*, Annual Review of Sociology 43:101-119
- [8] MILGRAM S., (1967), *The Small-World Problem*, Psychology Today 1:60-67

- [9] COLMENA J. S., (1958), *Relation analysis: the study of social organizations with survey methods*, Hum. Organ 17:28-36
- [10] GOODMAN L. A., (1961), *Snowball Sampling*, Ann. Matb. Stat. 32:148-70