Celem projektu badawczego jest określenie, jak struktura sieci społecznej i jakie charakterystyki programistów tworzących oprogramowanie Open Source sprzyjają powstawaniu innowacji w społeczności Open Source. Innowacje stanowią powstające w ramach współpracy pomiędzy programistami nowe programy lub poprawki.

Prowadzone w tym projekcie analizy oparte zostaną na unikatowej bazie zebranej z wykorzystaniem techniki *web scrapping* serwisu GitHub, stanowiącego obecnie jeden z największych serwisów związanych z oprogramowaniem Open Source. Baza zawiera publicznie dostępne dane umieszczone na stronach programistów zarejestrowanych w serwisie. Programiści wchodzą w różne relacje ze sobą (np. otrzymują informacje o działaniach innych osób, o ich projektach, czy współtworzą oprogramowanie), dlatego reprezentujemy ich jako elementy (węzły) sieci społecznej.

W oparciu o powyższą bazę sprawdzimy, czy reputacja programisty (mierzona m.in. za pomocą liczby osób, które wyraziły zgodę na otrzymywanie wiadomości o jego aktywności oraz liczbą przyznanych wyróżnień projektom programisty) zwiększa prawdopodobieństwo powstania innowacji w ramach współpracy z nim. Zbadamy też relację pomiędzy powstawaniem innowacji a zjawiskiem *homofilii* – czy współpracują ze sobą programiści o podobnych charakterystykach.

Pierwszym zadaniem będzie uporządkowanie i potencjalne uzupełnienie danych surowych pobieranych z serwisu GitHub i publicznych baz danych GHTorrent i GithubArchive. Dane dotyczące programistów są dostępne publicznie, jednak bardzo rozproszone i mogą zawierać zanieczyszczenia, które na tym etapie analiz powinny zostać usunięte.

Drugie zadanie obejmuje przygotowanie bazy danych w konkretnej postaci wymaganej dla modeli analitycznych wykorzystywanych w badaniu. Część danych jednostkowych wymaga ujednolicenia i agregacji. Dodatkowo, w tym zadaniu odbędzie się umieszczenie badanych sieci społecznych w przestrzeni hiperbolicznej. Wykorzystanie grafów hiperbolicznych do reprezentacji sieci społecznych jest stosunkowo nowym podejściem, pozwalającym opisać zależności w sieciach, które tworzą się zarówno dzięki zjawisku popularności danego węzła, jak i podobieństwa między węzłami.

W trzecim zadaniu sprawdzimy, jak struktura sieci, w której działają programiści wpływa na prawdopodobieństwo zaistnienia między nimi innowacji. Spróbujemy też określić, jakie charakterystyki programistów sprzyjają podjęciu z nimi współpracy.

Czwarte zadanie to walidacja wyników otrzymanych we wcześniejszych zadaniach. Na podstawie zebranej drugiej fali danych (obecnego stanu sieci społecznej), będziemy się starali przewidzieć, czy pomiędzy dwojgiem programistów pojawiła się innowacja w oparciu o narzędzia przygotowane we wcześniejszych zadaniach. Prognozowany graf porównamy z rzeczywistym. Ten krok pozwoli nam sprawdzić, na ile otrzymane przez nas ilościowe wyniki mogą być uogólniane.

Dotychczasowe badania społeczności Open Source dotyczyły zazwyczaj grup 100-300 programistów, skupionych w konkretnym projekcie. Praktycznie nie korzystano z analizy sieci społecznych oraz metod ekonometrycznych do ich analizy. Prezentowany projekt stara się wypełnić te luki badawcze. Badanie łączy tematycznie zarówno obszary informatyki, ekonomii, jak i innych nauk społecznych oraz wykorzystuje unikatową, reprezentatywną dla społeczności Open Source bazę danych.